Jabarian, Brian; Sartori, Elia

Working Paper

Critical Thinking and Storytelling Contexts

CESifo Working Paper, No. 11282

Provided in Cooperation with:
Ifo Institute – Leibniz Institute for Economic Research at the University of Munich

Suggested Citation: Jabarian, Brian; Sartori, Elia (2024) : Critical Thinking and Storytelling Contexts, CESifo Working Paper, No. 11282, CESifo GmbH, Munich

This Version is available at:
https://hdl.handle.net/10419/305524

CESifo WORKING PAPERS

11282
2024

August 2024

Critical Thinking and Storytelling Contexts

Brian Jabarian, Elia Sartori

CESifo

Impressum:

CESifo Working Papers
ISSN 2364-1428 (electronic version)
Publisher and distributor: Munich Society for the Promotion of Economic Research - CESifo GmbH
The international platform of Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute
Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740, Telefax +49 (0)89 2180-17845, email office@cesifo.de
Editor: Clemens Fuest
https://www.cesifo.org/en/wp
An electronic version of the paper may be downloaded
· from the SSRN website: www.SSRN.com
· from the RePEc website: www.RePEc.org
· from the CESifo website: https://www.cesifo.org/en/wp

CESifo *Working Paper No. 11282*

Critical Thinking and Storytelling Contexts

Abstract

We argue that storytelling contexts – the way information is communicated through varying credibility sources, visual designs, writing styles, and content delivery – impact the effectiveness of surveys and elections in eliciting preferences formed through critical thinking (reasoned preferences). Through an artefactual field experiment with a US sample (N = 725), incentivized by an (LLM), we find that intermediate storytelling contexts prompt critical thinking more effectively than basic or sophisticated ones. Sensitivity to these contexts is linked to individual cognitive traits, and participants with a high need for cognition are particularly responsive to intermediate contexts. In a conceptual framework, we explore how critical thinkers impact the efficiency of elections and polls in aggregating reasoned preferences. Storytelling contexts that effectively prompt critical thinking improve election efficiency. However, the in-decisiveness of critical thinkers can have ambiguous effects on election bias, potentially posing challenges for principals who are required to act on these election outcomes.

*Brian Jabarian**
Booth Business School, University of Chicago / IL / USA
brian.jabarian@chicagobooth.edu

Elia Sartori
Center for Studies in Economics and Finance and Università degli Studi di Napoli Federico II, Naples / Italy
elia.sartori@unina.it

*corresponding author

First version: March 29, 2023 - Current version: August 1, 2024
We are indebted to Roland Bénabou for his continued guidance. We are grateful for the comments of Erik Brynjolfsson, Colin Camerer, Tore Ellingsen, Nicolas Jacquemet, Luca Henkel, Yves Le Yaouanq, John List, Dan McGee, Muriel Nierdele, Pietro Ortoleva, Devin Pope, Eldar Shafir, Avner Strulov-Shlain, Jean-Marc Tallon, Richard Thaler, Marie Claire Villeval, George Wu, Leeat Yariv, and Sam Zbarsky. Rishane Dassanayake and Yeeun Koh provided exceptional research assistance, and Alfio De Angelis and Alessandro Sciacchetano provided excellent research assistance. We are grateful for the research participation of psychologists and colleagues from the Department of Psychology at Princeton University. We thank the seminar attendees at Stanford HAI, Google, Microsoft, Caltech, Chicago Booth, Chicago Econ, Bologna, and PSE. This paper was written in part while Brian visited the Department of Economics at Princeton University and the Kahneman-Treisman Center for Behavioral Science and Public Policy in 2018-2020. He thanks them for their hospitality. Ethics: We obtained Princeton IRB approval #12995. Funding: Brian acknowledges financial support from the Paris School of Economics, the Sorbonne Economics Center, and the Forethought Foundation for funding his visit to Princeton, Effective Venture US, Grant ANR-17-CE26-0003, and Grant ANR-17-EURE-001.

1 Introduction

Motivation. Individuals often hold preferences on complex issues involving persistent trade-offs between legitimate pros and cons (Kaplan, 1972). This complexity can lead individuals to form their preferences through different mental models (Tversky and Kahneman, 1974), using either an intuitive approach (System 1) or a reasoned approach (System 2) as summarized by Kahneman (2011). When relying on System 2, individuals engage in critical thinking, transitioning from automatic and naive responses to more analytical and reflective processing. This transition involves becoming aware of the inherent trade-offs in complex issues, enabling a more nuanced understanding and forming a reasoned preference (Halpern, 2013). Storytelling contexts – that is, the way facts are communicated through different writing styles, visual design, source credibility, and content delivery – play a crucial role in shaping behavior, as documented early on by cognitive psychologists (Tversky and Kahneman, 1981; Kahneman and Tversky, 1984) and, since then, extensively in behavioral, marketing, and media sciences (Pennycook et al., 2021; Kemp et al., 2022; Udry and Barber, 2024). It is then only natural to ask whether storytelling contexts also affect the formation of preferences in System 2, the transition to becoming aware of the complexity of an issue.

In this paper, we explore how principals can effectively target their actions based on these preferences formed through critical thinking, within relevant populations, spanning applications from political economy to industrial organization. While our conceptual framework will later detail here are briefly the two primary cases we consider. Overall, both cases hinge on the fact that standard mechanisms used to aggregate preferences does not capture the underlying mental model forming such preferences, leaving out relevant information for decision-makers. First, positive principals, like firms or social media influencers, depend on public approval for economic returns and social image. They use private polls to gauge reasoned preferences and minimize backlash when addressing controversial issues. Second, normative principals, such as social planners, navigate decisions on ambivalent issues where policy outcomes involve trade-offs between competing worldviews. They rely on the distribution of citizens' reasoned preferences as normative guidance for the "right" policy, ensuring voters understand trade-offs through critical thinking to reflect informed judgments.

Experimental Design and Results. Starting with our experimental framework, our artefactual field experiment on a US population ($N = 725$) is structured around three main components. Participants are initially classified as naive or critical thinkers through a combination of self-reported measures, open-ended responses, and an

incentivized knowledge test. Subsequently, they are exposed to one of three storytelling contexts, each presenting the same pros and cons regarding the issue at hand. These contexts varied in four dimensions: writing style (from slang to elaborated), user experience design (from rich to minimalist), source credibility (who and how many share the pros), and mode of content delivery (fragmented or continuous). Using LLMs, we generated these contexts, ensuring familiarity through three contexts (participants were made aware that contexts were artificially generated and facts were authentic): a simple Twitter-like context, an intermediate Facebook-like context, and a sophisticated newspaper-like context. The participants are then asked to write an essay, incentivized and assessed through LLMs, and compared against a US average score. To create a measure of critical thinking, we asked expert cognitive psychologists who evaluated essays, classifying submissions as pass or fail based on clear indications of critical thinking.

We find that an intermediate storytelling context (a medium-length social media environment) is more efficient in prompting participants to think critically than a simplistic context ($z = -2.171, p < 0.05$). Besides, we show that individuals' sensitivity to the storytelling format is linked to cognitive traits. In particular, participants with a high need for cognition (Cacioppo and Petty, 1982) are more sensitive to an intermediate storytelling context than in a simplistic context ($z = -3.001, p < 0.01$) and than a sophisticated context ($z = -2.168, p < 0.05$).

We conducted two types of robustness checks to evaluate our pre-treatment and post-treatment classification instruments. Our sensitivity analysis reveals that *Facebook* consistently outperforms *Twitter* and *Newspapers* in fostering critical thinking across various knowledge thresholds. At a high threshold (9 correct answers out of 10), *Facebook* shows significant superiority over *Twitter* ($z = -3.223; p < 0.01$) and *Newspapers* ($z = -2.451; p < 0.05$). At a lower threshold (7 out of 10), *Facebook* maintains its advantage ($z = -3.051; p < 0.01$). However, when the threshold drops to 5, the difference between *Facebook* and *Twitter* is less pronounced ($z = -2.674; p < 0.01$), supporting the notion that a baseline level of knowledge is essential for effective critical thinking, aligning with (Halpern, 2013).

We explored whether human expertise in grading critical thinking could be approximated by language features of the essay (e.g., length), or grades provided by our baseline LLM or expert ones. The length-based classification showed inconsistencies, as *Facebook* appeared less effective than *Twitter* ($z = 1.843; p < 0.1$) and *Newspapers* ($z = -0.171; p = 0.086$), with a moderate correlation between essay length and critical thinking ($r = 0.34$). In addition, psychologists were able to identify AI-generated essays with accuracy 59% to 92%, and there was a weak correlation ($r = 0.25$) between the number of AI essays evaluated and the accuracy of the identification. Further analysis showed that LLMs, such as GPT-3 and ChatGPT-4+,

cannot replicate human expertise in assessing critical thinking. GPT-3 scores did not show significant differences in thinking styles across storytelling contexts, like *Twitter* ($z = -0.755; p = 0.055$) and *Facebook* ($z = -1.326; p = 0.056$). Additionally, ChatGPT-4+ had a 44.87% mismatch rate with human evaluations, even with similar instructions. These results highlight the limitations of current LLMs and the ongoing necessity of evaluating human experts in this domain.

Conceptual Framework. In our model, the proportion of critical thinkers becomes a crucial welfare variable, affecting how accurately elections reflect the distribution of reasoned preferences in a large population. Principals can use election results to guide their decisions; therefore, it is crucial that the outcome of the election accurately predicts the distribution of stable preferences. Because we focus on a large (continuum) population, there are no issues of aggregation or strategic voting. However, citizens might not report their reasoned preferences (system 2) because they have not yet recognized the ambivalent nature of the issue (and vote instead based on instincts, system 1). Hence, the predictive power of reported preferences (election outcomes) regarding the distribution of reasoned preferences depends on the share of citizens that vote according to either system. We model critical thinking as "losing," a systematic voting tendency that is orthogonal to reasoned preferences. This abstraction compellingly captures the process of transitioning from an instinctive opinion to embracing an informed stance. For this reason, election outcomes typically become more informative about the average reasoned preference the larger the share of critical thinkers.

As mentioned above, eliciting preferences formed through critical thinking is relevant for two broad economic situations, corresponding to the principals' specifications. Consider the two previous cases in more detail. First, consider a positive principal – a private entity like a social media influencer or a firm under intense public scrutiny – whose social image or economic returns depend on the approval or rejection of a public stance on a controversial issue. They must make a statement about the issue that will act as a focusing event, meaning it will make the public discover the reasoned preference based on which they will judge the principal. To minimize backlash and maintain loyalty, the principal needs to understand the distribution of reasoned preferences within the public. A (private) poll can help uncover these preferences. However, there is a risk that respondents may express instinctive, naive opinions rather than those that will actually be used to judge the principal's statement.

Second, consider an institutional (or normative) principal, such as a social planner, who must choose a policy on an ambivalent issue. Even if they were omniscient—able to quantify the impact of any policy fully—the decision would still

hinge on selecting between two opposing worldviews. As one possible normative criterion, the planner might adopt the distribution of citizens' reasoned preferences, which, in this view, represents "the right thing to do." Such planners must then ensure that voters approach the polls with a thorough understanding of the consequences of their choices. For such a principal, each citizen is born with an equal portion of "truth" (his reasoned preference) but has to go through a critical thinking process to retrieve it fully.

For concreteness, suppose they need to decide on the size of a redistribution policy based on whether it is "right" to live in an egalitarian society or in a society that rewards each citizen in proportion to their contribution to societal well-being. While the principal might have a personal opinion on whether a free or egalitarian society is "better," they use the proportion of citizens who, after careful consideration, prefer an egalitarian society over a free one as a measure of "the right balance" between the two. Notice that her goal is to choose the most appropriate policy rather than focusing on re-election; in fact, the re-election issue would essentially represent a variation of the first type of principal, potentially constrained by the action limitations discussed later.

The distinction between the two types of principals in our analysis lies not only in the interpretation of their objectives but also in how they use surveys or election outcomes. Unlike surveys, which positive principals use solely for information, elections most likely determine the normative principal's action (e.g., by shaping the parliament or amending a law), leaving no room for adjustment based on the information obtained. As we will formalize in Section 5, this restriction drives qualitative differences in the impact of storytelling context on our welfare measure. Critical thinkers always increase the efficiency of elections, measured by the information the election outcome contains about the relevant unknown. However, they might introduce a bias in the election outcome that only the positive principle can correct, as such an outcome binds the normative one. For this reason, contexts that prompt citizens to think more quickly are unambiguously better for positive principles. For normative principals, however, the interaction with bias might lead, under some parameterizations, to perplexing comparative statics.

Contributions. Our paper directly contributes to the increasingly recognized role that *contexts*, beyond facts, play in shaping identity, preferences, and behaviors in economics (Kahneman, 2003; Bénabou and Tirole, 2016; Thaler, 2016; Bordalo et al., 2020) and in 'the cognitive turn in behavioral economics' (Enke, 2024). Building on the role of stories, whether understood as narratives (Bénabou and Tirole, 2002; Shiller, 2017; Bénabou et al., 2018), opinions (Bursztyn et al., 2022) or worldviews (Eliaz and Spiegler, 2020; Montiel Olea et al., 2022; Bordalo et al., 2023), we show

that two similar stories, because communicated through different storytelling contexts, will prompt different mental models, upon which preferences are formed and behaviors are shaped. We focus, in particular, on one mental model, critical thinking, and contribute to operationalizing the pyramidal framework proposed by List (2022). This framework decomposes Kahneman's "System 2" into four multidimensional thinking levels, ranging from naive thinking ("modal"), passing between intermediate levels ("neophyte" and "adept") to the highest level ("great"). Our method enables us to partially identify the transition from level 1 to level 4, focusing on the dimensions related to attention (Schwartzstein, 2014; Caplin, 2016; Loewenstein and Wojtowicz, 2024), memory (Bordalo et al., 2021) and salience (Bordalo et al., 2022): becoming aware of a specific trait of a complex issue, that is, the fact that the issue is a trade-off without hard conclusions. In this sense, our paper contributes to defining complexity beyond the domain of lotteries (Oprea, 2020; Banovetz and Oprea, 2023; Kendall and Oprea, 2024), showing that a feature of complexity can be understood through forming a mental model (Kendall and Oprea, 2024) and relying on procedural decision-making (Banovetz and Oprea, 2023; Arrieta and Nielsen, 2024) to make difficult decisions when facts are absent (Halevy et al., 2023).

To establish such results, our experiment developed different techniques. It contributes to the very recent use of open-ended responses to measure mental models (Stantcheva and Ferrario, 2022; Haaland et al., 2024) and, to our knowledge, is the first to show how to use LLMs to ensure the internal validity of storytelling contexts and to incentivize the identification of critical thinking. Related to AI techniques, our paper also contributes to documenting the potential of AI personalization beyond standard demographics (Rafieian and Yoganarasimhan, 2023). Since participants with a high need for cognition drive our key result, principals could optimally manipulate such a finding to expose agents in the 'right' storytelling context, matching their cognitive styles.

Finally, a key reason why the behavioral literature has increasingly focused on mental states and models lies in their implications for economic policy and welfare (Bernheim and Rangel, 2007; Bernheim et al., 2021, 2024). Specifically, our conceptual framework elaborates on the cases where a principal is incentivized to elicit and aggregate agents' reasoned preferences to maximize welfare when relying on surveys, as increasingly done in behavioral macroeconomics and economic policy (Kuziemko et al., 2015; Stantcheva, 2021; Binetti et al., 2024) and normative welfare in public and political settings relying on voting and elections (Feddersen and Pesendorfer, 1997; Kim and Fey, 2007; Bhattacharya, 2013). Furthermore, we find that critical thinkers are more likely to form not only reasoned preferences but also stable ones ($\chi^2 = 5.12$, $p = 0.0236$), which highlights a relevant feature of criti-

cal thinking as a behavioral mechanism potentially relevant to minimize cognitive noise or uncertainty (Enke and Graeber, 2023) and finally, to reduce measurement error in contexts where standard techniques requiring double elicitation such as ORIV (Gillen et al., 2019) are not feasible due to the choice environment like voting where voting only once for a given election is allowed. Finally, based on our conceptual framework, List et al. (2024)'s recent findings on the reduction of misinformation spread thanks to critical thinking can be interpreted and tracked as a welfare-maximizing policy objective.

Structure. The remainder of the paper is structured as follows. Section 2 details the experimental design. Section 3 details our main experimental results. Section 4 discusses the robustness checks. Section 5 describes the behavioral welfare model and its main positive and normative results. Section 6 discusses the connection between our experimental and conceptual findings, suggesting respective extensions. Section 7 concludes.

2 Experimental Design

In this section, we provide a brief overview of the experiment, discuss the pre- and post- treatment strategies to classify participants as critical thinker vs naive thinker (summarized in Table 1), how we measured preferences and cognitive styles and detail our incentive mechanism procedure.

In our experiment, we expose subject participants to different storytelling contexts and elicit their mental models in the critical thinking process, pre-and post-treatment.[1] We then test whether the likelihood of transitioning from naive thinkers to critical thinkers varies significantly between contexts. Throughout the experiment, we also collected data about participants' cognitive styles. We used incentivized elicitations for key individual variables pre and post-critical thinking mental models and implemented diverse algorithms to ensure data collection quality. Figure 1 provides an overview of the experimental design and its primary elicitations, which we elaborate on in subsequent sections.

[1] Full details of our data collection process can be found in Appendix B.1.

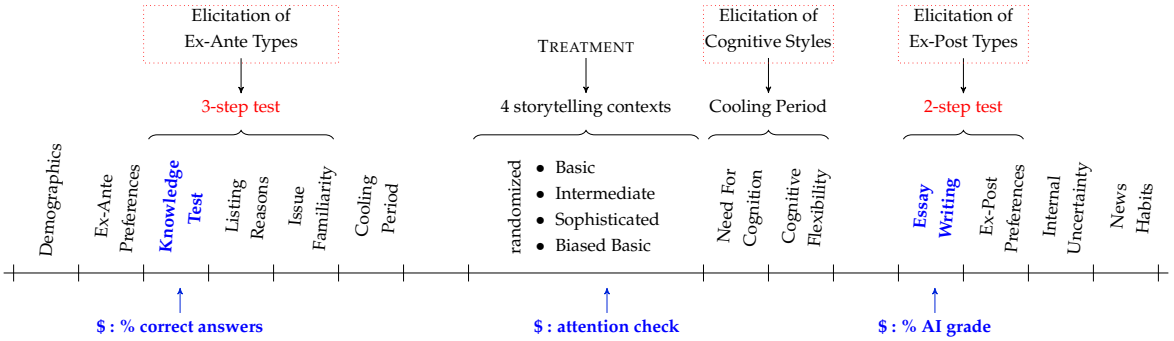


Figure 1: Experiment Design

2.1 Pre-Treatment Classification

To be classified as a critical thinker before treatment, participants must satisfy the following three conditions: i) *accuracy*: having basic knowledge of the complex issue; ii) *engagement*: having already thought deeply about the complex issue; iii) *listing*: being capable of listing reasons for being pros and cons of the complex issue. Failing one of these conditions, participants are classified as *naive thinker*.

The three conditions are complementary to each other to minimize the risks of misclassification. First, *accuracy* is necessary to be a critical thinker: evaluating a complex issue without basic knowledge is a non-starter (Halpern, 2013). Second, *Engagement* is a useful complement to *accuracy* to ensure robustness: a participant satisfying *engagement* but not *accuracy* will be finally classified as a non-critical thinker. Third, *Listing* immediately refers to the *neophyte thinking* state elaborated by List (2022): a participant capable of reasoning starts to show a transition from System 1 to System 2, but not enough to be considered a *critical thinker* because of potential memory or mimetic effects in cases where conditions *accuracy* and *engagement* are respectively not met (Kahneman, 2011).

To generate the *accuracy* condition, we rely on an incentivized standard Pew Research knowledge test on digital privacy (Vogels and Anderson, 2019). The test contains ten questions related to digital privacy. To generate the *engagement* condition, participants self-reported whether they had thought deeply about the issue before coming to our experiment. To generate the *listing* condition, participants must provide two reasons that support their preference for the issue of digital privacy and two that are against it.

2.2 Storytelling Contexts Treatments

After the pre-treatment classification, participants are randomly assigned to one of the four treatments. These treatments keep the facts constant but vary in the different storytelling contexts in which they are communicated.

Such storytelling contexts differ in four dimensions of user experience (UX) / user interface (UI) design: writing style, visual design, source credibility, and content delivery. First, writing style refers to the vocabulary, tone, length of sentences, and coherence transitions between them. Such a style varies from crude to moderate and finally to sophisticated. Second, visual design refers to layout, color schemes, and typography. Such a design varies from maximalist to minimalist. Third, source credibility in the storytelling context refers to the perceived trustworthiness and reliability of the information source. Source credibility varies from multiple anonymous individuals sharing different individual stories to a single institutional instance aggregating all the different individual stories within a coherent and long-aggregated story. Fourth, content delivery in the storytelling context refers to how stories are structured and presented to the audience. Content delivery varies from displaying stories one at a time on separate screens (one story per screen frame) to presenting a continuous, long-form story all within a single screen frame.

Having the ecological validity of our experiment in mind, which could quickly become artificial and lose its intended potential effects, it is crucial to design realistic experimental environments, an approach increasingly favored in economics (Harrison and List, 2004), not shifting the standard in natural field experiments but also in lab experiments (Mol, 2019) and artefactual experiments (Innocenti, 2017; Andries et al., 2024). It is even more crucial if one, as we, aims to document the role of *contexts* on mental models, preferences, and behaviors.

In our case, where digital familiarity is the main condition for the 'natural context', we generated three main storytelling contexts familiar to the US population as documented in recent Pew Research and Gallup US survey reports (pew, 2020; gal, 2020): *Twitter*, *Facebook*, and *Newspaper*. These platforms were chosen because they represent distinct variations in writing style, visual design, source credibility, and content delivery, allowing us to explore how each dimension influences user perception and engagement. Importantly, the total length, total number of the same facts, and duration of interaction with each fact are kept constant in the different contexts.

In all treatments, tweets, Facebook posts, and news articles arrive in a random order sequentially (one per screen) and remain on screen for a given fixed amount of time (participants cannot move to the next screen by themselves). This ensures

that the total length, total number of facts, and duration to interact with each fact remain consistent across different contexts. All treatments have a duration of 120 seconds, as summarized in Table 4 in Appendix, except for *Biased Twitter*, which lasts half of this duration due to its one-sided structure.

To generate such storytelling contexts, we employed different neural networks and LLM techniques (level GPT3) available at the time of data collection.[2] Participants were explicitly informed before the treatments and at the end of experiment that the tweets, Facebook posts were not real to ensure no deception was involved.

2.3 Post-Treatment Classification

After being exposed to the different storytelling treatments, we again classify the mental models of the participants to identify possible transitions from naive thinking to critical thinking. We cannot rely on the same three-pronged tests from the pre-treatment classification since the condition *Listing* storytelling context treatments expose subjects to a series of pros and cons about the issue. Relying on the same condition here can misidentify critical thinking as memory effects since individuals with better memory would be better at recalling information obtained from treatments.

Alternatively, we ask participants to write a critical thinking essay on the issue at hand. We then rely on the expertise of Princeton's Ph.D.-level and above cognitive psychologists, specializing in critical thinking and closely related cognitive psychology topics, to identify whether participants are critical thinkers or not. Let us elaborate on our measure in more detail. After the storytelling context treatment, participants are reminded of their pre-treatment preferences on the issue. Then, they are asked to write a critical thinking essay.

Participants are informed that their essays will be graded by an LLM, Grammarly, trained on millions of data points on the US population, and focused on measuring 'writing quality.' Although this LLM is efficient in assessing the overall quality of writing, it cannot capture the nuances of critical thinking, especially in terms of discerning whether the writer demonstrates an awareness of the ambivalence surrounding the issue. To address this limitation, we ask cognitive psychologists with Ph.D. degrees to provide a professional assessment of the essays. These experts are randomly and independently assigned to the participants' essays. They are asked to evaluate whether the essay reflects a state of awareness or not, assigning a pass or fail grade accordingly. While participants receive payment based on the AI's evaluation, our analysis focuses on the cognitive psychologists' assessment, with AI's scores serving as a robustness check.

[2]See Section B.3 in Appendix for more details.

We recruited 15 psychologists (doctoral level or above) who specialize in cognitive psychology at Princeton University. Each grader was randomly assigned a "grading treatment" (that is, a set of essays to grade). This set of essays was randomly constructed and consists of essays from the four treatments. In addition, the graders were not informed about the treatment to which subjects were assigned. Psychologists must grade a very short paragraph as follows. The grading consists of giving a passing grade if psychologists judge that the participant "realizes that the issue is ambivalent," a failing grade otherwise. What may happen is to confound high cognitive sophistication (i.e., the ability to write well-written essays in English), facilitated by the fact that they read some arguments right before this essay exercise with their self-reasoning skill "realizing that the issue is ambivalent", which is the variable that we want to elicit. This is a specific case that is still challenging for AI-based grading software and is the main reason why human expertise is uniquely useful.

Each grader was paid a fixed fee of $50 for each grade session. Each grader could participate up to three times in our experiment, and no grader could be assigned twice to the same grading treatment. For robustness, each essay was corrected three times by different psychologists. Despite "triple-eliciting" such grades, this metric can still be prone to measurement error. Hence, the estimated intensity *levels* should be used with caution. However, our focus is on treatment differences, which remain reliable since the psychologists were unaware of each subject's specific treatment.

Here is the summary of our pre- and post-treatment classification strategies:

Treatment	C	N
BEFORE	Knowledge Test Score $> \tau_{KTS}$ Issue Familiarity = 1 Reasons List Both Side = 1	Else
AFTER	Psychologists Grade = Pass	Else

Table 1: Classification Strategy before/after treatment

2.4 Preferences and Cognitive Styles

Ex-ante and ex-post preferences. Before the treatment, we prompt participants on different political issues (i.e., without a baseline): guns, crime, climate, welfare, and digital privacy issues. We use the standard congressional metrics, including digital issues. We elicit more than only digital preferences to ensure that participants do not guess at this stage which preferences we focus on in the remainder of the experiment (treatment and critical thinking essay) to minimize their social desirability

bias. After the treatment on digital privacy, we survey participants again to elicit their preferences about digital privacy.[3]

Cognitive styles. Throughout the experiment, we measured the cognitive styles of the participants to explore whether cognitive personalization was necessary, that is, personalizing the exposure to specific storytelling contexts conditional on cognitive styles matters to change from naive thinking to critical thinking. We rely on two well-established metrics to measure cognitive styles: the Need for Cognition Scale (NCS) and the Cognitive Flexibility Scale (CFS).

The former scale, NCS, measures a participant's propensity to think deeply. Neuroscientists and cognitive psychologists proposed it, and it has become standard in cognitive psychology (Cacioppo and Petty, 1982).[4] It comprises a series of six questions that each receive a score between 1 and 5. Examples of questions include "I prefer to think about small, daily projects to long-term ones" and "Learning new ways to think doesn't excite me very much". We compare the aggregate score with the sample average to classify participants into high or low need for cognition.

The latter, CFS, is a standard metric in cognitive psychology proposed by Martin and Rubin (1995), measuring an agent's ability to switch between thoughts and courses of action. It comprises a series of six questions that each receive a score between 1 and 6. We compare the aggregate score with the average of the US population to classify the participants as having high or low cognitive flexibility.

2.5 Incentive Mechanism and Quality Checks

In the experiment, participants receive two types of payment. First, they receive a fixed reward of $2 for fully completing the experiment by answering the comprehension questions correctly, guaranteed. Second, they receive a bonus payment of up to $6. This bonus payment itself comes from two incentivized tasks: the knowledge test before treatment, where they can earn up to $1, and the essay task after treatment, where they can earn up to $5. We ask participants to write two short essays during this study that will be graded from 0 to 100 points using an LLM. We divide the bonus payment into two parts.

The largest part (from $0 to $5) is proportional to the weighted average score on the essay writing task; the second essay receives more weight (2/3) because it requires more writing (400 characters as opposed to 200 characters). The score can range from 0 to 100 points, and the reward will be proportional to the score. If the

[3] For a comprehensive list of our elicitation measures, see Appendix B.5.

[4] See Cacioppo et al. (1996); Furnham and Thorne (2013); Cacioppo et al. (1983); Wu et al. (2014); Lord and Putrevu (2006); Lins de Holanda Coelho et al. (2020) as examples of studies that make use of and develop the NCS.

participants score 0, they win \$0. If they get a score of 50, they win \$2.50. If they get a score of 100, then they win \$5. An essay that receives a low AI score can still earn a high score for critical thinking and awareness. This accounts for any penalization that may be administered to the essays of participants who are less fluent in English. In the instructions to psychologist graders, we define and exemplify what we mean by a "dilemma," "realizing that the issue is ambivalent," and "critical thinking". We also run robustness checks with philosophers.

To be eligible for the remaining bonus payment (up to \$1), participants must receive at least an average score of 50/100 in the essay exercise in addition to the bonus of the writing essay. This requirement ensures that participants take the exercise seriously; cheaters and agents who are inconsistent in their preferences are not eligible for this bonus payment. The performance of the participant on the knowledge test determines this additional bonus. The test consists of 10 questions, and each participant receives \$0.10 for each correctly answered question.

We implement three attention screeners, which are standard in online experimental economics. The core of our experiment is for participants to write an original essay by themselves. As such, we need the subjects to avoid accessing external information during the writing task. We implement two algorithms to monitor cheating behavior. Before starting the experiment, we informed participants that they must not access external information during the experiment, particularly during the knowledge test and the essay exercise. Failing to do so would be considered "cheating behavior". As such, they would be red-flagged and prevented from receiving the additional bonus payment of \$1.

3 Main Results

We gathered 860 participants from a representative US population using Prolific, a data collection platform increasingly favored by economists due to its high data quality. After attention and quality screening, our final sample size was $N = 725$.[5] Participants received a fixed payment of \$2 and a bonus payment of up to \$5, resulting in an average payment of approximately \$6.

3.1 Storytelling Contexts Matter for Prompting Critical Thinking

To explore the role of storytelling contexts in the critical thinking process of individuals, we calculate, for each treatment $i \in \{Newspaper, Twitter, Facebook, Biased\ Twitter\}$, the frequency $\hat{\lambda}_i$ with which agents subject to the storytelling context i transition from naive thinking to the critical thinking. We then used these intensities to

[5] See Section XX in Appendix for more details.

perform a difference-in-mean test of the null hypotheses $\lambda_i = \lambda_j$ for all possible combinations of treatments $\{i, j\}$.[6]

We find that the only significant difference is between *Facebook* and *Twitter*, where the former performs better when transitioning subjects from a thinking style N to C ($z = -2.171, p < 0.05$) as detailed in Table 2.

Treatment	*Newspaper*	*Twitter*	Biased *Twitter*	*Facebook*
Newspaper	·	·	·	·
Twitter	1.141 (0.049)	·	·	·
Biased *Twitter*	−0.272 (0.052)	−1.440 (0.049)	·	·
Facebook	−0.967 (0.052)	−2.171** (0.049)	−0.702 (0.052)	·
N	170	190	184	181

Standard errors in parentheses
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 2: *Z-Score* Difference-in-Proportions

We hypothesize that Twitter's format encourages more shallow engagement with content compared to Facebook, thus impeding the transition into critical thinking. Twitter's character limit forces users to condense complex ideas into brief statements, often oversimplifying the content.[7] This brevity can restrict the depth of information, limiting the ability of users to engage critically. Tversky and Kahneman (1974) supports this idea by highlighting how simplified information can lead to heuristic processing rather than systematic analysis.

Additionally, Twitter's fast-paced flow of information emphasizes real-time updates, which can discourage users from pausing to reflect and analyze what they consume. The constant influx of new posts creates an environment where users are continuously bombarded with stimuli, reducing opportunities for deep reflection. These structural features of Twitter emphasize short, attention-grabbing headlines and sound bites, which further contribute to surface-level engagement. Pennycook et al. (2021) further supports this idea by providing evidence that deliberately drawing attention to the accuracy of a news story on Twitter reduces the spread of misinformation, thereby supporting the claim that, by default, Twitter's fast-paced nature reduces the scope for critical thinking. Furthermore, Bago et al. (2020) shows that participants who were placed under time pressure were more likely to believe false headlines than those who were allowed to deliberate (b = 0.36, 95% CI = [0.2, 0.52], p < 0.0001). This implies that faster consumption of news may inhibit the careful,

[6]Formally, $\hat{\lambda}_i = \frac{\#(N \to C)_i}{\#(N \to C)_i + \#(N \to N)_i}$, where N refers to naive thinking and C to critical thinking.

[7]This experiment was designed and launched before Musk's Twitter era, which led to the increase in tweet lengths for Blue Twitter users.

deliberate cognition associated with critical thinking.

Moreover, while Facebook encourages in-depth and personal connections, Twitter emphasizes brief and anonymized exchanges. Oz et al. (2018) shows, using a quantitative content analysis of 1,458 posts responding to White House social media accounts, that Twitter posts exhibited less deliberative attitudes compared to*Facebook*(F(1, 1457) = 204.52, p < .001). They also found that comments posted on *Twitter* were significantly more uncivil (F(1, 1457) = 110.55, p < .001) and impolite (F(1, 1457) = 33.83, p < .001). This contrast in user engagement may contribute to the observed differences in the effectiveness of storytelling contexts in promoting critical thinking.

Our results extend the existing literature on the impact of social media platforms as storytelling contexts on user behavior and cognition. Mena (2020) and Figl et al. (2023) demonstrate the effectiveness of warning labels and certificates in reducing the sharing of false news on Facebook. Murthy et al. (2015) show that tweeting on mobile phones encourages more immediate egocentric content. Munson et al. (2013) and Rieger et al. (2023) explore web browser-based interventions to mitigate selective exposure and confirmation bias. We add that variations in storytelling contexts between platforms, specifically *Twitter* and Facebook, affect the degree to which users are prompted to think critically. Despite our discussion above of possible mechanisms driving this relationship, more research is needed.

3.2 Cognitive Styles Drive the Efficiency of Storytelling Contexts

We explore how individuals' cognitive styles affect the magnitude of storytelling's effects. We find that an individual with a high need for cognition is more likely to transition to critical thinking when exposed to *Facebook* storytelling contexts compared to *Twitter* and *Newspaper* types (respectively, $z = -3.001, p < 0.01$ and $z = -2.168, p < 0.05$). Compared to Table 2, Table 3 delineates statistically significant differences in proportions between *Facebook* and *Twitter*. In our sample, 42% of 725 participants identified as having a high need for cognition, and 33% identified as having high cognitive flexibility.

Treatment	*Newspaper*	*Twitter*	*Biased Twitter*	*Facebook*
Newspaper	.	.	.	.
Twitter	0.744	.	.	.
	(0.072)			
Biased Twitter	−1.340	−2.149**	.	.
	(0.079)	(0.074)		
Facebook	−2.168**	−3.001***	−0.843	.
	(0.081)	(0.076)	(0.083)	
N	71	77	81	72

Standard errors in parentheses
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 3: *z-score* FOR *High Need for Cognition.*

This difference highlights the role of cognitive styles in driving the efficiency of storytelling contexts to prompt critical thinking. It also documents how much cognitive personalization should be done carefully to avoid inefficient matching and fully exploit its potential (Rafieian and Yoganarasimhan, 2023). In particular, based on these findings, a principal can avoid exposing a high need for cognition type with a sophisticated storytelling context. In addition, we do not find any effect on cognitive flexibility. Relatedly, Mosleh et al. (2021) finds that Twitter users with higher scores on the Cognitive Reflection Task (CRT) (Frederick, 2005) are more likely to share content from reliable news sources and use words associated with insight (OR = 1.138, p < 0.001) and inhibition (OR = 1.133, p < 0.001), indicating a more deliberative cognitive style. Our findings extend this by showing that individuals with a high need for cognition are more effectively prompted to think critically through Facebook's detailed posts than Twitter's concise format. This suggests that the effectiveness of storytelling contexts in fostering critical thinking is significantly influenced by the cognitive style of the individual, particularly their need for cognition.

4 Robustness Checks

We performed two types of robustness checks related to our pre-treatment and post-treatment classification instruments.

4.1 Pre-Treatment Classification Robustness

Our sensitivity analysis shows that our results remain robust regardless of the thresholds used to classify the mental models of participants prior to treatment, focus-

ing on both the digital knowledge test and the reason list exercise.[8] Participants needed to score at least 7 out of 10 on the digital privacy knowledge test, a significantly more demanding threshold than the Pew Research standard, where only 20% scored seven or more. Our sample had an average score of 7.146, with a median of 8. Despite this higher threshold, we evaluated whether the effectiveness of treatment depended on the score above or below 7.

Facebook consistently fosters more critical thinking than *Twitter* or *Newspaper* for individuals with a high need for cognition. For example, at a knowledge threshold of 10, *Facebook* outperforms *Twitter* ($z = -2.968, p < 0.01$) and *Newspaper* ($z = -2.438, p < 0.05$). Similarly, at a threshold of 9, the advantage of *Facebook* over *Twitter* ($z = -3.223, p < 0.01$) and *Newspaper* ($z = -2.451, p < 0.05$) is statistically significant.[9] Even when the threshold is reduced to 5, *Facebook* maintains a notable advantage over *Twitter* ($z = -2.674, p < 0.01$). However, significance decreases compared to higher thresholds, highlighting the importance of sufficient baseline knowledge to assess critical thinking engagement accurately.

Facebook's effectiveness in promoting critical thinking is further validated when considering knowledge scores above 8, where it remains more impactful than *Twitter* ($z = -3.051, p < 0.01$) and *Newspaper* ($z = -2.229, p < 0.05$). These findings persist in stringent settings, demonstrating *Facebook*'s effectiveness in prompting critical thinking across various knowledge levels. When the knowledge test threshold is set at 8 with a requirement to list three reasons, *Facebook* still shows greater efficacy compared to *Twitter* ($z = -3.051, p < 0.01$) and *Newspaper* ($z = -2.229, p < 0.05$). In particular, we find that for participants with higher knowledge scores, *Facebook*'s impact remains pronounced compared to *Twitter* ($z = -2.122, p < 0.05$) and *Biased Twitter* ($z = -1.013, p = 0.049$).

Finally, as the threshold decreases, the statistical significance between *Facebook* and *Twitter* decreases ($z = -0.458, p = 0.054$), reinforcing the idea that adequate knowledge is crucial to the effectiveness of the media platform in fostering critical thinking.[10] This pattern aligns with the general trend observed, where higher knowledge thresholds correlate with stronger treatment effects.

4.2 Post-Treatment Classification Robustness

This subsection examines the reliability of human experts versus LLM agents in grading critical thinking essays. We assess the limitations of using essay length as a proxy for critical thinking, explore the accuracy of psychologists in identifying AI-generated essays, and compare the effectiveness of GPT-3 and ChatGPT-4+ in

[8] See Section C.1 in Appendix for more details.
[9] See Table 9 in Appendix for more details.
[10] See Table 10 in Appendix for more details.

evaluating critical thinking. Our findings highlight the complexity of accurately assessing critical thinking, underscoring the necessity of human judgment over automated metrics.

Length-Based Metrics to Detect Critical Thinking Are Misleading. Besides, we investigate whether essay length can effectively replace human psychologists' grading as a measure of critical thinking. Our findings demonstrate that human judgment is essential, as essay length alone does not reliably indicate critical thinking skills. Specifically, *Facebook* essays were less effective in promoting critical thinking compared to *Twitter* ($z = 1.843; p < 0.1$) and *Newspaper* ($z = -0.171; p = 0.086$), contradicting length-based assessments that underestimated Facebook's impact.

Although there is a moderate positive correlation between essay length and critical thinking ($r = 0.34$), length alone is insufficient for comprehensive evaluation. Additionally, psychologists' accuracy in identifying AI-generated essays ranged from 59% to 92%, with a weak positive correlation ($r = 0.25$) between the number of AI essays evaluated and identification accuracy.[11]

Aligning LLMs with Human Experts We asked psychologists whether they believed that the essays they had to grade were written by an AI or a human. We found that the accuracy rates were highly variable, ranging from 59% to 92%, with a mean accuracy rate of 76.9%. Additionally, we find a weak positive correlation ($\rho = 0.24$) between the number of AI essays evaluated and the accuracy rate, suggesting that exposure to AI essays offers slight improvement.

Based on this previous weak correlation, we wondered to which extent human expertise and feedback data are necessary for the grading system. We explore whether LLMs can replace human expert grading in evaluating critical thinking. In general, the analysis confirms that current LLMs lack the depth and accuracy needed for critical thinking assessments, emphasizing the continued need for human expertise in this domain.[12]

On the one hand, the baseline LLM level GPT-3 (Grammarly) was tested to determine its effectiveness in identifying critical thinking within essays. The analysis did not reveal significant differences in changing thinking styles across storytelling contexts like *Twitter* ($z = -0.755; p = 0.055$) and *Facebook* ($z = -1.326; p = 0.056$). This indicates that GPT-3 cannot accurately detect critical thinking signals. On the other hand, the more advanced *ChatGPT-4+* was evaluated for its ability to detect critical thinking. The results showed a 44.87% mismatch between the LLM and

[11] See Tables 13 and 14 in Appendix for more details.

[12] Below, we summarize the results of our four different robustness checks and direct the reader to Tables 15, 16, 18, 17 and 18 in Appendix for more details on each.

psychologist grades. Even with similar instructions, 48.01% of the grades did not align. This highlights the limitations of ChatGPT-4+ in replicating human evaluation, especially in contexts like *Twitter* ($z = -1.567; p = 0.086$) and *Facebook* ($z = -1.720; p = 0.086$).

On the other hand, essay quality, including word count and grammar, was found to correlate with higher critical thinking assessments. Higher word counts ($\beta = 0.001; p < 0.001$) and better grammar ($\beta = 0.005; p < 0.001$) increased the likelihood of passing grades. LLMs may not fully capture these confounding factors. Finally, for participants with a high need for cognition, LLM grades did not show significant differences in thinking styles between storytelling contexts, suggesting that LLMs cannot effectively replace human grading for this subgroup.

5 Conceptual Framework

In this section, we develop a simple model in which the intensity of the critical thinking process affects the efficiency of preference aggregation through surveys (in industrial organization settings) or elections (in political economy settings). In our stylized social choice setting, welfare is the distance between the political action and a target determined by the distribution of reasoned preferences, i.e., those held after completing the *critical thinking* process, within the population.

This parameter is unknown at first and can only be estimated using the result of a poll held at some time t, when (a part of) the citizens may still not have completed their process. The mechanics of our model are relatively straightforward: as more citizens become critical thinkers, the election outcome becomes less dependent on a component of preferences orthogonal to the reasoned preference – the relevant unknown. A faster critical thinking process increases the share of critical thinkers and enhances the efficiency of elections.

This section proceeds as follows. We first present the two welfare benchmarks (corresponding to the two types of principals discussed in the introduction) and the critical thinking process separately. Combining the two, we then obtain closed-form for the evolution of welfare and establish under what conditions a faster critical thinking process is beneficial. The proofs of the main results and some immediate extensions are relegated to the appendix.

5.1 Two Welfare Benchmarks

The relevant unknown is the distribution of reasoned preferences in a large population over a binary alternative. This is the share $p \in [0, 1]$ of agents who would prefer outcome 1 to outcome 0 if they had completed a critical thinking process.

Welfare realizes the distance between an action a – taken by the principal – and its target p:[13]

$$W(a,p) = -(a-p)^2$$

Ex-ante, p is unknown and drawn from a normal distribution $p \sim \mathcal{N}(\mu,\sigma)$. Therefore, without a poll to elicit the population's preferences, the principal would choose $a = \mu$ and obtain the value $-\sigma^2$.[14] Before choosing $a \in [0,1]$, the principal observes the proportion $\bar{p}$ of agents that report preferring the alternative 1. We call $\bar{p}$ the *election outcome*. Consider two types of principals that differ in the use they can make of this information.

Positive Principal. A positive principal, P, for which the election outcome is *not* binding, namely, who can choose any $a \in [0,1]$ regardless of the implementation of $\bar{p}$. The positive principal uses the election result and his knowledge of the critical thinking process within the population to estimate p. His optimal action is the conditional expectation

$$a^\star = \hat{p} := \mathbb{E}[p|\bar{p}]$$

that achieves value;

$$W_P = -\mathbb{E}\left[(\hat{p}-p)^2\right], \tag{1}$$

equal to the dispersion of the conditional mean $\hat{p}$ around p. Both expectation operators $\mathbb{E}$ integrate under the joint distribution of $p, \bar{p}, \hat{p}$, which depend on the voting behavior and citizens' critical thinking process — which we derive in the next section.

Connecting to the discussion in the introduction, one can think of such principals as public figures (e.g., multinational firms or social influencers with reputational concerns) who have to take a stance on a trade-off. They privately run a poll and use its outcome as they wish to fine-tune their statement. Payoff depends on the (distribution of) reasoned preferences because the statement acts as a "focusing

[13] Although the space of reasoned preferences — individuals' resolution of the trade-off — is binary, the policy space is continuous. This corresponds to a situation where the planner can fine-tune the policy to the *distribution* of individuals' reasoned preferences. Think, for example, of a social planner choosing the size of the welfare program based on the share of agents who hold an egalitarian (rather than a free market) view. A different specification with $a^\star = \mathbb{I}[p > \frac{1}{2}]$ (binary action space) provides similar insights but is less tractable.

[14] The normality assumption gives tractable conditional expectations and closed-form welfare. Obviously, it is inconsistent with the compact support $[0,1]$. The analysis with ex-ante uniform p (and p_S) is algebraically more involved but does not change the qualitative results. For the sake of tractability, we keep the normal setup, implicitly assuming that σ is "small enough" that the mass outside $[0,1]$ is negligible.

event" that pushes the relevant population into critical thinking: the preferences individuals judge the principal on are (potentially) different from those they report at the poll.

Institutional Principal. Second, we consider an institutional principal, I, who has to choose $a = \bar{p}$. One can think of such principal as democratic institutions that must comply with the election outcome (say, by empowering a parliament whose composition is proportional to $\bar{p}$).[15] Due to the constraint in their action, I achieves value.

$$W_I = -\mathbb{E}\left[(\bar{p} - p)^2\right] \tag{2}$$

via the standard decomposition, we obtain

$$W_I = W_P - B, \tag{3}$$

where

$$B = \mathbb{E}\left[(\bar{p} - \hat{p})^2\right] > 0$$

It is the bias of election, representing how the average reported preference differs systematically from the reasoned preferences. A principal P who can correct for such social tendencies only suffers from the dispersion of the estimator $\hat{p}$ around the parameter p, while the principal I must also be concerned with the bias of the election.

5.2 Thinking Styles and Voting Processes

We now turn to describing the voting behavior of individuals (or citizens). Each citizen is characterized by a reasoned (or stable) preference

$$y \sim Ber(p)$$

Where p is the unknown welfare relevant to which the principal wants to match. However, if asked in a poll, citizens do not necessarily report their reasoned preference. This is because the reasoned preference is "discovered" at the end of a *critical thinking process* that individuals undergo.

Critical Thinking. Citizens transition through two critical thinking states $\{N, C\}$, where N means *Naive* and C means *Critical Thinking*. We assume that the critical

[15] In this context, the interpretation of p differs. Rather than focusing on the potential backlash from reasoned preferences, we envision an institutional principal considering p as a normative criterion for aggregating social preferences about a dilemma. Essentially, the distribution of preferences of individuals who have undergone the critical thinking process determines the "right thing to do".

thinking process follows a simple dynamic in continuous time: all individuals start at $t = 0$ in the state N and, independently of y (and other voting parameters), transition to the absorbing state C with intensity $\lambda \in (0, \infty)$. Therefore, at time t, there will be a fraction

$$\eta_N = \exp\{-\lambda t\}$$

of agents that are still *N*aive and $\eta_C = 1 - \eta_N$ that transitioned to be *C*ritical thinkers.[16] The parameter λ is key for our analysis. It represents the intensity with which individuals realize that the issue at hand is ambivalent. In our experiment, we established that the way the fact is presented (storytelling context) has an effect on λ and that this effect depends on the cognitive abilities of the population. Notice that in models where the principal cares exclusively about the share of critical thinkers, the result is trivial since η_C increases in λ for all t.

Voting Behavior. We denote x the preference that individuals report in the polls and assume that it depends on the reasoned preference y and on the stage of the critical thinking process $\{N, C\}$. Before realizing that the issue is ambivalent, the preference reported x_N is

$$x_N | y = \begin{cases} Ber(p_N) & \text{w.p. } 1 - \beta \\ y & \text{w.p. } \beta \end{cases}$$

In other words, x_N is equal to the reasoned preference with probability $\beta \in [0, 1]$, while the complementary probability is drawn from a distribution of naive preferences $p_N \sim \mathcal{N}(\mu, \sigma)$, independent of p. Since we still have a parameter β driving the correlation between average stereotypes and reasoned preferences, the assumption of independence is harmless. It only requires the formation of naive preferences involving factors not solely related to p.[17]

Notice that a high β represents situations where, despite not realizing the ambivalent nature of the issue, stereotypes get their reasoned preference right with a high probability. On the contrary, $\beta = 0$, corresponding to a situation where stereotypes are independent of reasoned preferences, an election held at $t = 0$ (all stereotypes) would yield $\bar{p} = p_N$, completely uninformative of p.

The preference reported by individuals in *C* loses its dependence on the nui-

[16] The assumption that *C* is an absorbing state, with no switch from *C* to *N*, captures the idea that awareness is an irreversible process. A straightforward extension of the model prevents a scenario in which all individuals eventually reach the state *C*: a constant fraction $\nu < \lambda$ exits the economy and reenters in the awareness state *N*. Qualitative results would remain unchanged as the associated share of naive agents: $\eta_N(t) = \frac{\nu}{\lambda} + \exp(-\lambda t)\left(1 - \frac{\nu}{\lambda}\right)$ would still be decreasing in λ,t.

[17] The identical distribution of p,p_N is instead for tractability alone. Most of the derivations in the appendix utilize non-identically distributed normal variables (μ, σ, μ_N, σ_N). We discuss such extensions, focusing on the meaning of $\mu \neq \mu_N$, in Section 6.1.

sance parameter p_N and becomes a function of the reasoned preference alone,

$$x_C \,|y = \begin{cases} y & \text{w.p. } \xi \\ 1-y & \text{w.p. } 1-\xi \end{cases}$$

The parameter $\xi \in [\frac{1}{2}, 1]$ represents in reduced form situations in which citizens realize that the issue is ambivalent but have not yet found their reasoned preference. The case $\xi = 1$ corresponds to a situation in which individuals discover their reasoned preference immediately after realizing the ambivalence of the issue; $\xi = \frac{1}{2}$ represents instead the opposite situation of permanent indecisiveness of C individuals. We think of our two-stage critical thinking process as a reduced form of a fully identified three-stage process – detailed in the Appendix – where C is an intermediate stage where agents have realized the ambivalent nature of the issue but have not formed their reasoned preference yet, i.e., they are in a phase of normative uncertainty. Note that if all individuals are in the state C (that is, a poll held at $t \to \infty$), then the election result is $\bar{p} = \xi \cdot p + (1-\xi) \cdot (1-p)$, which is a strictly monotonic (hence invertible) function of p if $\xi > \frac{1}{2}$. In that case, $\hat{p} = p$, and the positive principal is chosen efficiently. For interior shares η, the election outcome is given by:

$$\overline{p} = \eta_N \left(\beta p_N + (1-\beta)\, p\right) + \eta_C \left(\xi \cdot p + (1-\xi) \cdot (1-p)\right)) \tag{4}$$

And the parameter ξ also affects positive welfare. We can now use the (joint) normality assumption to write $\bar{p}$ and the conditional expectation $\hat{p}$ as a linear function of the fundamental unknowns p, p_N, that is,

$$\bar{p} = \alpha_0 + \alpha_1 \cdot p + \alpha_2 \cdot p_N$$

$$\hat{p} = \gamma_0 + \gamma_1 \cdot p + \gamma_2 \cdot p_N$$

where loadings $\boldsymbol{\alpha}, \boldsymbol{\gamma}$ are functions of the structural parameters $\vartheta = [\beta, \xi, \boldsymbol{\mu}, \boldsymbol{\sigma}]$ and the statistic of the critical thinking process η (see the appendix). Once we specify the joint normal expectation operator, we can compute (the evolution of) both positive and institutional welfare (2)-(3) in closed form and arrive at our main result.

Proposition 1 *i) For all values of structural parameters ϑ, W_P is increasing in t and λ.*

ii) W^I has nontrivial comparative statics in λ, t. If $\beta < 1-\xi$, then it is monotonically increasing; if $\beta > \frac{(1-\xi)\left((1-2\mu)^2+4\sigma^2\right)}{2\sigma^2}$ then it is monotonically decreasing; else it grows locally to $t = 0$ (resp. $\lambda = 0$) up to a finite time $t^\star$ (finite intensity $\lambda^\star$) then eventually decreases.

We guide our discussion of the results presented in the proposition by follow-

ing the graphical representations in Figure 2, Panels A-D. First, notice the Positive Welfare (blue line) is increasing in time in all the represented cases — indeed, point $i)$ establishes that this is true for all values of ϑ: if the principal can leverage the election outcomes without constraints, then a faster transition to critical thinking (higher λ) leads to more efficient elections. This occurs because, as fewer individuals are naive, elections allow for a more efficient estimation of welfare-relevant parameter p as they are less affected by the noise in p_N. This finding is significant as it establishes λ as a welfare measure: storytelling contexts that accelerate the shift to critical thinking (see Panel 2d) are unequivocally preferred by positive principals.

In point (ii), we highlight a potential limitation of this result for principals who must act based on the election outcome (Institutional welfare, yellow lines). Since they have to "play" the election result, these principals must consider the (potentially adverse) effect that the shift towards critical thinking has on election bias.

The key parameter determining how λ (or, equivalently, t) impacts the institutional welfare is β, namely how predictive the preferences of naive citizens are of their stable preferences. If β is low, welfare increases (Panel 2a); for intermediate β, welfare initially increases, reaches an interior maximum, and then declines (Panel 2b); for high β, welfare decreases monotonically (Panel 2c). Although this may seem paradoxical, it has a natural explanation: when β is high, naive preferences strongly predict reasoned preferences (in the extreme case where $\beta = 1$, naive citizens vote according to their stable preferences despite not recognizing the ambivalent nature of the issue). Therefore, the transition to critical thinking introduces an attenuation bias due to the indecisiveness ξ of critical thinkers.

In sum, the parameters β, ξ capture the inherent quality of the preference reported in the two stages of the critical thinking process. High-β (resp, high ξ) environments represent situations where naive citizens (resp, critical thinkers) get their reasoned preference right with high probability. As high λ facilitates the transition from naive to critical thinkers, it tends to improve efficiency when the latter are relatively more reliable. Importantly, however, only naive voters have a component in their reported preference that is unknown to the principal and cannot be filtered out. Therefore, the positive principal always benefits from higher λ, no matter the values of β, ξ— though the effect will be larger if β is low relative to ξ. The institutional principal, instead, cannot correct the bias. Therefore, the sizes of β, ξ affect whether critical thinkers are desirable at all: the pool of critical thinkers contains more information about p but might report it less accurately. For intermediate values of β, the counteracting forces give rise to hump-shaped welfare with an interior optimal share of critical thinkers.

A natural question of potential independent interest concerns the evolution of election bias. Indeed, the principal might aim to reduce this bias by designing the

intensity of the critical thinking process λ (or choosing the election time) so that the election outcome is an unbiased estimator of p, potentially sacrificing some efficiency. To address this question, we investigate conditions under which the loadings $\boldsymbol{\alpha}, \gamma$ coincide "by divine coincidence," ensuring that the positive and institutional principals have the same action rule and hence the same value.

By imposing $\boldsymbol{\alpha} = \gamma$ we identify a share of stereotypes $\eta^\star$ where the two coincide. Consequently, there exists an intermediate time at which the average reported preference is unbiased for p. Formalizing this result, we obtain the following:

Proposition 2 *If $\beta > \frac{1}{2}$, then there exists a finite time $t^\star$ such that $B(t^\star) = 0$. If, in addition $\xi = 1$, $t^\star$ is given by*

$$t^\star = -\frac{1}{\lambda} \log\left(\frac{1}{2\beta}\right),$$

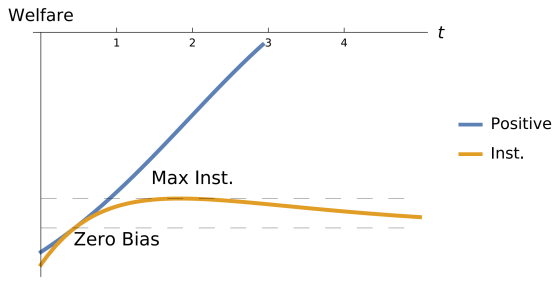
with immediate comparative statics.

Hence, an institutional principal, I, can use the timing of elections (equivalently, the speed of the critical thinking process) to remove the bias of the election. However, this approach does not achieve the welfare-maximizing outcome. We demonstrate that Institutional welfare increases with the share of critical thinkers around the zero-bias level $\eta^\star$. This implies that the principal is willing to tolerate a slight bias in order to enhance efficiency further. A graphical illustration is given in Panel 2b of Figure 2: the welfare-maximizing time is situated to the right of the zero-bias time.

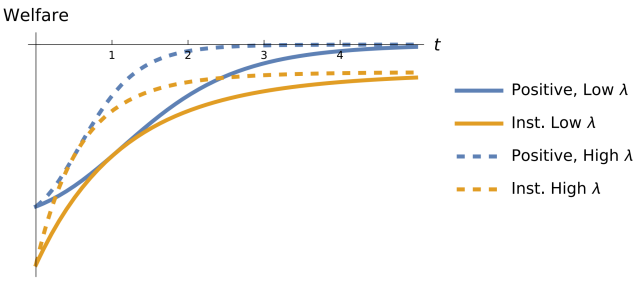
6 Discussion and Extensions

In Section 5, we have presented a relatively parsimonious model of voting while undergoing a critical thinking process (from naive to critical thinkers). Our analysis shows that the effectiveness in promoting critical thinking, which varies across storytelling contexts, as experimentally confirmed in Section 3, plays a crucial role in determining the accuracy of polls. In this section, we elaborate on the connection between our empirical findings and conceptual propositions to highlight our main conclusions and discuss potential extensions of our work to address related research questions.

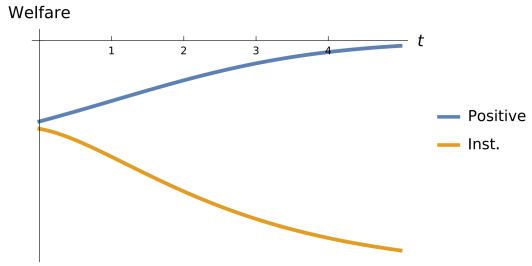
We have established that the intensity λ of the critical thinking process – what we estimated for different storytelling contexts – is a relevant welfare measure: Regardless of the time of the elections, a higher intensity increases the information content of the elections – combining this result with the results of our experiment



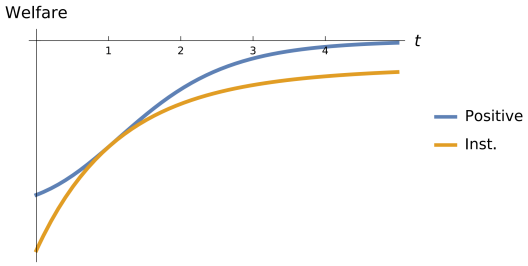
(a) Intermediate β: Interior maximum (above zero-bias)



(b) Welfare with low (solid) and high (dashed) λ.



(c) High β: decreasing W^I



(d) Low $\beta < 1 - \xi$: increasing W^I

Figure 2: Evolution of Positive (blue) and Institutional Welfare in time (horizontal axis) for parametrizations that give different cases according to Proposition 1.

(Section 3) indicates that the way information is communicated to citizens prior to voting affects the quality of information elicited in a poll.

We also show that the unambiguous comparative statics only holds if the principal can freely manipulate the results of the poll when taking their action (which we refer to as the positive principal, P). If the outcome of the election constrained their action, as is most likely the case for an institutional principal, I, principal, then a bias-precision trade-off makes the comparative statistics ambiguous: A faster critical thinking process might hurt the efficiency of elections (Propositions 1-*ii*) and 2).

Importantly, in our model, there is no intrinsic social value for having many critical thinkers. This is somehow restrictive, as "mental flexibility" could have social benefits beyond increasing election accuracy (see, e.g., Bernheim et al. (2021)). Most likely, this direct effect is more relevant for an Institutional type of principal. In this sense, our model underestimates the welfare effect of storytelling contexts as it disregards the potential direct benefits of having a larger share of critical thinkers.

6.1 Conceptual Discussion and Extensions

Addressing preference aggregation in the presence of an electorate segment that has not yet recognized the ambivalent nature of the problem is a significant question. Our model takes a first step by providing a conceptual foundation and presenting

preliminary results. However, a comprehensive analysis should address potential asymmetries and incorporate adjustment margins based on empirical evidence.

In particular, the strongest assumption we implicitly maintain is that of *symmetry*: the voting parameters β, ξ are independent of reasoned preferences (or naive) preference. A natural relaxation of this assumption is to posit that.

$$\beta_i = \mathbb{P}\left[x_S = y | y = i\right]$$

and

$$\xi_i = \mathbb{P}\left[x_A = y | y = i\right]$$

with a different specification for residual uncertainty in naive preferences.[18] Insofar as overconfidence can be interpreted as individuals' resistance to critical thinking, evidence in Ortoleva and Snowberg (2015) also questions the fact that intensity λ is independent of y: if the reasoned preference predicts cognitive traits associated with critical thinking (or the impact of different storytelling contexts), then the pool of critical thinkers would be selected based on y, which constitutes an additional source of bias.

Another potentially relevant relaxation of symmetry is to allow for the presence of *bias in naive preferences*, which means to allow for $\mu_N \neq \mu$. This corresponds to a situation where the principal knows that a specific opinion is prevalent before individuals realize the ambivalent nature of an issue. This scenario seems particularly relevant when one position is more likely to be defended through superficial arguments such as nationalism. If this is the case, then an institutional principal would benefit from increasing the intensity λ (or simply "letting time pass") as critical thinkers would naturally eliminate this systematic bias.[19]

Extending the model to account for these asymmetries would not alter our main conclusion: storytelling is crucial for election efficiency because polls yield more information about the distribution of reasoned preferences when there are more critical thinkers. However, these asymmetries could significantly influence the voting *behavior of critical thinkers* who have not yet discovered their stable preferences. We did not model this form of indecisiveness directly, but instead captured it in a reduced form through the ξ parameter. "Early-stage" critical thinkers, having lost their naive preferences but not yet developed their reasoned preferences, experience a state of "normative uncertainty" (MacAskill et al., 2020; Millner, 2020). In

[18]The symmetry hypothesis can readily be tested in experiments like ours where we observe individuals before starting their critical thinking process and after discovering their stable preference – i.e., using the three mental models extension of the model –. In our specific setting, we could perform such a test due to a lack of power (the number of subjects who were classified as reasoned preferences at the end of the experiment was small).

[19]Obviously, the evolution of welfare for the positive principal would instead be unaffected by this extension, as their could "clear out" all systematic noise in the poll.

this state, they might attempt to counterbalance the bias present in the electorate due to asymmetries in voter types. For example, they might vote for the option that they perceive is underrepresented among stereotypes (e.g., 0 if $\mu_N > \mu$) or because those who hold such stereotypes transition to critical thinking more rapidly.

Moreover, due to their indecisiveness, such voters are probably more likely to abstain and are more sensitive to *voting costs*. Hence, our analysis suggests that voting costs might not just select the electorate based only on demographic heterogeneity (see (Cantoni and Pons, 2022)) but potentially also on the critical thinking stage.

Lastly, discounting the utility of delaying action allows us to explore the *optimal timing of elections* within our framework. Even the positive principal, who chooses efficiently in the limit, would not delay their decision indefinitely under discounting. Analyzing how the optimal timing of elections varies with λ and other parameters involves examining how a principal can control the type and duration of storytelling to maximize the accuracy of polls. Such issues are relevant in many practical scenarios, such as designing the type and accuracy of information provided to a focus group before soliciting their opinions on a marketing campaign or policy proposal.

Overall, the simple model introduced and analyzed in this paper already sheds light on an issue relevant to public and private entities that need to elicit the distribution of reasoned preferences within a population. Our analysis suggests that these entities should consider and possibly control the information environment in which the population learns about an issue, as well as the cognitive traits prevalent in that population. For example, an influencer aiming to gain approval from their audience should carefully consider how an issue is presented and discussed. Similarly, a government planning a referendum on an ethically sensitive issue should be mindful of the media's presentation and discussion of the issue. This government, which must accept the election outcome at face value and cannot manipulate it to extract the 'best predictor', should also consider additional bias margins that could skew results and create counterintuitive comparative statics.

6.2 Experimental Discussion and Extensions

Internal validity. The internal validity of our critical thinking measure can be further strengthened by using advanced monitoring algorithms with a larger online sample. Specifically, we propose clustering participant behavior by analyzing whether they open new browser tabs or take significant breaks during the essay task. This could indicate accessing external information through another tab or device. This monitoring can help identify instances where participants seek external

information, which can affect their critical thinking responses. Additionally, analyzing keystroke patterns and character counts can detect live plagiarism, a concern that becomes increasingly relevant as LLMs evolve and become more sophisticated, rendering traditional detection methods less effective. By addressing these factors, we aim to bolster the reliability and accuracy of our measure.

To enhance the robustness of human expertise in grading critical thinking essays, it is crucial to consider factors beyond a mere sample size expansion of experts. Cognitive psychologists' traits, such as institutional affiliations, subfields, political views, and expertise quality (e.g., publications and citations), can introduce biases in grading, especially if evaluators exhibit predispositions toward certain pros or cons stories, say, because of the participants taking a stance matching the graders' political preferences. Leveraging techniques such as Reinforcement Learning from Human Feedback (RLHF) can help calibrate grading systems by aggregating evaluations from diverse expert opinions, thereby mitigating measurement biases and ensuring an even more balanced assessment of critical thinking.

Incentive mechanisms can also be optimized using economic principles, moving beyond traditional flat payments for graded batches. Designing incentive structures that align with desired grading outcomes can enhance accuracy and reliability, encouraging more consistent and unbiased assessments such as beauty contest-based, or Bayesian truth serum-based incentives could be relevant. Furthermore, the interdisciplinary nature of critical thinking invites collaboration with cross-field experts, including educators, philosophers, and scholars from human sciences, to enrich the grading process with diverse perspectives and methodologies, ultimately leading to a more holistic evaluation framework.

External Validity. The external validity of our experiment must be interpreted with caution. Our study should not be viewed merely as a comparison of social networks, such as *Facebook* versus Twitter. Instead, it explores how storytelling contexts influence critical thinking. The key takeaway is that "the storytelling context matters," rather than concluding that "Facebook is better." Social media, in this study, serves as a storytelling medium where individuals are exposed to various perspectives, often superficial. The critical question is whether this exposure aids individuals in recognizing the trade-offs involved in complex issues or whether it requires more in-depth study and personal reflection. While such an approach fosters reasoned preferences, it is time-consuming and less likely to occur naturally.

At the end of our experiment, we asked several questions related to participants' habits to consume political news: frequency, which types of media and which social media platforms. Our findings suggest that for *daily* users from our sample seeking political information but we did not find any evidence that the type of media or

the choice of social media platform does not impact the effects. Yet, we believe that more fine-grained news consumption habits research is warranted, as well as different personal traits that could impact the effectiveness of a storytelling context. *Facebook* may provide slightly more engaging or informative content compared to *Newspaper* ($z = 0.222, p < 0.1$) but *Facebook* clearly outperforms *Twitter* ($z = 0.375, p < 0.01$). Besides, we find similar results for the high-need-for-cognition participants: *Facebook* slightly outperforms *Newspaper* ($z = 0.371, p = 0.068$), indicating a slight advantage for *Facebook* and clearly outperforms *Twitter* ($z = -0.526, p = 0.006$).[20]

Our experimental design offers a foundation for exploring critical thinking across novel cross-context issues. While this study focuses on digital privacy, the storytelling features inherent to different issues may vary in their effectiveness at prompting critical thinking. Recognizing that issues possess an "objective" side that requires information from multiple sources, we emphasize the inherent trade-offs present in many issues. Critical thinking plays a pivotal role in guiding individuals to acknowledge the ambivalent nature of these issues. Our analysis highlights a novel channel, possibly orthogonal to previously identified ones, where the same characteristics that make hard information challenging to digest (e.g., continuous display with superficial language) may enhance attention to specific issues and foster awareness of their ambivalent nature.

A key area for future research is understanding the economic relevance of critical thinking in natural settings. Exploring how critical thinking emerges in complex environments, where agents on social media platforms encounter a mix of topics rather than clear-cut pro and con discussions, could provide insights into the effectiveness of storytelling contexts. In addition, natural settings offer opportunities to study the persistent effects of storytelling on critical thinking beyond the short-term timelines typically examined in A/B tests. These settings allow for empirical exploration of critical thinking's impact on outcomes such as voting behavior, as theorized in our framework. Understanding these dynamics could offer valuable implications for policymakers and educators aiming to foster critical thinking in society.

7 Conclusion

In this paper, we argue that the way information is communicated (storytelling context) may significantly influence the effectiveness of surveys and elections in eliciting reasoned preferences from a population. Our argument unfolds in two parts: first, we present empirical evidence from an incentivized experiment demonstrat-

[20] See Tables 11,12 in Appendix for more details on the news consumption habits analysis.

ing that storytelling context affects critical thinking. Second, we construct a conceptual framework that shows how critical thinking affects electoral efficiency.

Our empirical results document that the degree to which individuals become critical thinkers — recognizing the ambivalent feature of an issue, a prerequisite for forming their reasoned preferences — varies depending on the storytelling context in which otherwise identical information is communicated. Furthermore, we show that individuals' sensitivity to the storytelling format is linked to a cognitive trait (need for cognition) established in the cognitive psychology literature. The experiment assesses the critical thinking status of participants after different treatments through professional psychological expert evaluations. It employs an LLM to incentivize the writing task and diverse algorithms to detect cheating behavior and ensure data quality. We find that intermediate-length storytelling contexts (Facebook style) outperform others, particularly among individuals with a high need for cognition.

Then, in our conceptual framework, the proportion of critical thinkers is a key welfare variable, influencing how accurately elections reflect the distribution of stable preferences in a large population. Critical thinking is modeled as "losing" a systematic voting tendency, which is orthogonal to reasoned preferences and reduces the predictive power of reported preferences regarding the mean of reasoned preferences. This abstraction compellingly captures the process of transitioning from an instinctive opinion to embracing an informed ethical stance. However, the model's significant simplifications, particularly its assumption of symmetry in voting behavior and preference discovery based on instinctive opinions, may limit its applicability beyond the primary point of this paper, illustrating how and why critical thinking matters.

References

(2020), "American views 2020: Trust, media and democracy." URL https://knightfoundation.org/reports/american-views-2020-trust-media-and-democracy/.

(2020), "News use across social media platforms in 2020." URL https://www.journalism.org/wp-content/uploads/sites/8/2021/01/PJ_2021.01.12_News-and-Social-Media_FINAL.pdf.

Andries, Marianne, Leonardo Bursztyn, Thomas Chaney, and Milena Djourelova (2024), "In their shoes." Technical report, National Bureau of Economic Research.

Arrieta, Gonzalo and Kirby Nielsen (2024), "Procedural decision-making in the face of complexity." Technical report, Working Paper.

Bago, Bence, David G. Rand, and Gordon Pennycook (2020), "Fake news, fast and slow: Deliberation reduces belief in false (but not true) news headlines." *Journal of Experimental Psychology: General*.

Banovetz, James and Ryan Oprea (2023), "Complexity and procedural choice." *American Economic Journal: Microeconomics*, 15, 384–413.

Bénabou, Roland and Jean Tirole (2002), "Self-confidence and personal motivation." *The Quarterly Journal of Economics*, 117, 871–915.

Bénabou, Roland and Jean Tirole (2016), "Mindful economics: The production, consumption, and value of beliefs." *Journal of Economic Perspectives*, 30, 141–164.

Bernheim, B Douglas, Luca Braghieri, Alejandro Martínez-Marquina, and David Zuckerman (2021), "A theory of chosen preferences." *American Economic Review*, 111, 720–754.

Bernheim, B Douglas, Kristy Kim, and Dmitry Taubinsky (2024), "Welfare and the act of choosing." Technical report, National Bureau of Economic Research.

Bernheim, B Douglas and Antonio Rangel (2007), "Toward choice-theoretic foundations for behavioral welfare economics." *American Economic Review*, 97, 464–470.

Bhattacharya, Sourav (2013), "Preference monotonicity and information aggregation in elections." *Econometrica*, 81, 1229–1247.

Binetti, Alberto, Francesco Nuzzi, and Stefanie Stantcheva (2024), "People's understanding of inflation." Technical report, National Bureau of Economic Research.

Bordalo, Pedro, Katherine Coffman, Nicola Gennaioli, Frederik Schwerter, and Andrei Shleifer (2021), "Memory and representativeness." *Psychological Review*, 128, 71.

Bordalo, Pedro, John J Conlon, Nicola Gennaioli, Spencer Yongwook Kwon, and Andrei Shleifer (2023), "How people use statistics." Technical report, National Bureau of Economic Research.

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer (2020), "Memory, attention, and choice." *The Quarterly journal of economics*, 135, 1399–1442.

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer (2022), "Salience." *Annual Review of Economics*, 14, 521–544.

Bursztyn, Leonardo, Aakaash Rao, Christopher Roth, and David Yanagizawa-Drott (2022), "Opinions as facts." Technical Report 159, ECONtribute Discussion Paper.

Bénabou, Roland, Armin Falk, and Jean Tirole (2018), "Narratives, imperatives, and moral reasoning." Working Paper 24798, National Bureau of Economic Research, URL http://www.nber.org/papers/w24798.

Cacioppo, J. T., R. E. Petty, and K. J. Morris (1983), "Effects of need for cognition on message evaluation, recall, and persuasion." *Journal of Personality and Social Psychology*, 45, 805–818, URL https://doi.org/10.1037/0022-3514.45.4.805.

Cacioppo, John T and Richard E Petty (1982), "The need for cognition." *Journal of Personality and Social Psychology*, 42, 116.

Cacioppo, John T., Richard E. Petty, Jeffrey A. Feinstein, and William B. G. Jarvis (1996), "Dispositional differences in cognitive motivation: The life and times of individuals varying in need for cognition." *Psychological Bulletin*, 119, 197–253.

Cantoni, Enrico and Vincent Pons (2022), "Does context outweigh individual characteristics in driving voting behavior? evidence from relocations within the united states." *American Economic Review*, 112, 1226–1272.

Caplin, Andrew (2016), "Measuring and modeling attention." *Annual Review of Economics*, 8, 379–403.

Eliaz, Kfir and Ran Spiegler (2020), "A model of competing narratives." *American Economic Review*, 110, 3786–3816.

Enke, Benjamin (2024), "The cognitive turn in behavioral economics."

Enke, Benjamin and Thomas Graeber (2023), "Cognitive uncertainty." *The Quarterly Journal of Economics*, 138, 2021–2067.

Feddersen, Timothy and Wolfgang Pesendorfer (1997), "Voting behavior and information aggregation in elections with private information." *Econometrica: Journal of the Econometric Society*, 1029–1058.

Figl, K., Samuel Kießling, and Ulrich Remus (2023), "Do symbol and device matter? the effects of symbol choice of fake news flags and device on human interaction with fake news on social media platforms." *Computers in Human Behavior*.

Frederick, Shane (2005), "Cognitive reflection and decision making." *Journal of Economic Perspectives*, 19, 25–42, URL https://www.aeaweb.org/articles?id=10.1257/089533005775196732.

Furnham, Adrian and Jeremy D. Thorne (2013), "Need for cognition." *Journal of Individual Differences*, 34, 230–240, URL https://doi.org/10.1027/1614-0001/a000119.

Gillen, Ben, Erik Snowberg, and Leeat Yariv (2019), "Experimenting with measurement error: Techniques with applications to the caltech cohort study." *Journal of Political Economy*, 127, 1826–1863.

Haaland, Ingar K, Christopher Roth, Stefanie Stantcheva, and Johannes Wohlfart (2024), "Measuring what is top of mind." Technical report, National Bureau of Economic Research.

Halevy, Yoram, David Walker-Jones, and Lanny Zrill (2023), *Difficult decisions*. University of Toronto, Department of Economics.

Halpern, Diane F (2013), *Thought and Knowledge: An Introduction to Critical Thinking*. Psychology Press.

Harrison, Glenn W and John A List (2004), "Field experiments." *Journal of Economic literature*, 42, 1009–1055.

Innocenti, Alessandro (2017), "Virtual reality experiments in economics." *Journal of behavioral and experimental economics*, 69, 71–77.

Kahneman, Daniel (2003), "Maps of bounded rationality: Psychology for behavioral economics." *American economic review*, 93, 1449–1475.

Kahneman, Daniel (2011), *Thinking, Fast and Slow*. Macmillan.

Kahneman, Daniel and Amos Tversky (1984), "Choices, values, and frames." *American Psychologist*, 39, 341–350.

Kaplan, Kalman J (1972), "On the ambivalence-indifference problem in attitude theory and measurement: A suggested modification of the semantic differential technique." *Psychological Bulletin*, 77, 361–372.

Kemp, P. L., V. M. Loaiza, and C. N. Wahlheim (2022), "Fake news reminders and veracity labels differentially benefit memory and belief accuracy for news headlines." *Scientific Reports*, 12, 21829, URL https://doi.org/10.1038/s41598-022-25649-6.

Kendall, Chad and Ryan Oprea (2024), "On the complexity of forming mental models." *Quantitative Economics*, 15, 175–211.

Kim, Jaehoon and Mark Fey (2007), "The swing voter's curse with adversarial preferences." *Journal of Economic Theory*, 135, 236–252.

Kuziemko, Ilyana, Michael I Norton, Emmanuel Saez, and Stefanie Stantcheva (2015), "How elastic are preferences for redistribution? evidence from randomized survey experiments." *American Economic Review*, 105, 1478–1508.

Lins de Holanda Coelho, G., P. H. P. Hanel, and L. Wolf (2020), "The very efficient assessment of need for cognition: Developing a six-item version." *Assessment*, 27, 1870–1885, URL https://doi.org/10.1177/1073191118793208.

List, John A (2022), "Enhancing critical thinking skill formation: Getting fast thinkers to slow down." *The Journal of Economic Education*, 53, 100–108.

List, John A, Lina M Ramírez, Julia Seither, Jaime Unda, and Beatriz Vallejo (2024), "Toward an understanding of the economics of misinformation: Evidence from a demand side field experiment on critical thinking." Technical report, National Bureau of Economic Research.

Loewenstein, George and Zachary Wojtowicz (2024), "The economics of attention." *SSRN Electronic Journal*. Available at SSRN: https://www.ssrn.com.

Lord, K. R. and S. Putrevu (2006), "Exploring the dimensionality of the need for cognition scale." *Psychology & Marketing*, 23, 11–34, URL https://doi.org/10.1002/mar.20108.

MacAskill, Michael, Krister Bykvist, and Toby Ord (2020), *Moral uncertainty*. Oxford University Press.

Martin, Matthew M and Rebecca B Rubin (1995), "A new measure of cognitive flexibility." *Psychological Reports*, 76, 623–626.

Mena, P. (2020), "Cleaning up social media: The effect of warning labels on likelihood of sharing false news on facebook." *Policy & Internet*.

Millner, Anthony (2020), "Nondogmatic social discounting." *American Economic Review*, 110, 760–775.

Mol, Jantsje M (2019), "Goggles in the lab: Economic experiments in immersive virtual environments." *Journal of Behavioral and Experimental Economics*, 79, 155–164.

Montiel Olea, José Luis, Pietro Ortoleva, Mallesh M Pai, and Andrea Prat (2022), "Competing models." *The Quarterly Journal of Economics*, 137, 2419–2457.

Mosleh, Mohsen, Gordon Pennycook, Antonio A. Arechar, and David G. Rand (2021), "Cognitive reflection correlates with behavior on twitter." *Nature Communications*.

Munson, Sean A, Stephanie Lee, and P. Resnick (2013), "Encouraging reading of diverse political viewpoints with a browser widget." *ICWSM*.

Murthy, D., Sawyer Bowman, Alexander Gross, and Marisa McGarry (2015), "Do we tweet differently from our mobile devices? a study of language differences on mobile and web-based twitter platforms."

Oprea, Ryan (2020), "What makes a rule complex?" *American economic review*, 110, 3913–3951.

Ortoleva, Pietro and Erik Snowberg (2015), "Overconfidence in political behavior." *American Economic Review*, 105, 504–35.

Oz, M., P. Zheng, and G. M. Chen (2018), "Twitter versus facebook: Comparing incivility, impoliteness, and deliberative attributes." *New Media & Society*, 20, 3400–3419, URL https://doi.org/10.1177/1461444817749516.

Pennycook, G., Z. Epstein, M. Mosleh, and et al. (2021), "Shifting attention to accuracy can reduce misinformation online." *Nature*, 592, 590–595, URL https://doi.org/10.1038/s41586-021-03344-2. Received 05 March 2020; Accepted 08 February 2021; Published 17 March 2021; Issue Date 22 April 2021.

Rafieian, O. and H. Yoganarasimhan (2023), "Ai and personalization." In *Artificial Intelligence in Marketing (Review of Marketing Research, Vol. 20)* (K. Sudhir and O. Toubia, eds.), 77–102, Emerald Publishing Limited, Leeds.

Rieger, Alisa, Tim Draws, M. Theune, and N. Tintarev (2023), "Nudges to mitigate confirmation bias during web search on debated topics: Support vs. manipulation." *ACM Transactions on the Web*.

Schwartzstein, Joshua (2014), "Selective attention and learning." *Journal of the European Economic Association*, 12, 1423–1452, URL https://doi.org/10.1111/jeea.12104.

Shiller, Robert J (2017), "Narrative economics." *American Economic Review*, 107, 967–1004.

Stantcheva, Stefanie (2021), "Understanding tax policy: How do people reason?" *The Quarterly Journal of Economics*, 136, 2309–2369.

Stantcheva, Stefanie and Beatrice Ferrario (2022), "Eliciting people's first-order concerns: Text analysis of open-ended survey questions." *American Economic Association Papers and Proceedings (forthcoming)*.

Thaler, Richard H (2016), "Behavioral economics: Past, present, and future." *American economic review*, 106, 1577–1600.

Tversky, Amos and Daniel Kahneman (1974), "Judgment under uncertainty: Heuristics and biases." *Science*, 185, 1124–1131.

Tversky, Amos and Daniel Kahneman (1981), "The framing of decisions and the psychology of choice." *Science*, 211, 453–458.

Udry, J. and S. J. Barber (2024), "The illusory truth effect: A review of how repetition increases belief in misinformation." *Current Opinion in Psychology*, 56, 101736, URL https://doi.org/10.1016/j.copsyc.2023.101736.

Vogels, Emily A and Monica Anderson (2019), "Americans and digital knowledge." *Pew Research Center*.

Wu, C.-H., S. K. Parker, and J. P. J. de Jong (2014), "Need for cognition as an antecedent of individual innovation behavior." *Journal of Management*, 40, 1511–1534, URL https://doi.org/10.1177/0149206311429862.

Appendices

A Proofs of The Main Model

A.1 Preliminary Results on $\hat{p}$ and alike

The following three steps explicitly show how to analyze the evolution of the mental model process over time for each agent and how this relates to the parameters of the model.

1) $\mu_S = \exp\{-\lambda_1 t\}$ and $\mu_C = 1 - \mu_S$ Represent the masses. λ_1 represents the intensity with which agents transition from mental model N to mental model C over time. Moreover, we define the unknown parameter $\bar{p}$ as a function of μ and p

$$\begin{aligned}\overline{p}(\mu, p) &= \mu_S(\mathbb{E}[x_S|p]) + \mu_C(\mathbb{E}[x_C|p]) \\ &= \mu_S(\beta p_S + (1-\beta)p) + \mu_C(\xi_C p + (1-\xi_C)(1-p)) \\ &= \mu_S(\beta p_S + (1-\beta)p) + \mu_C(1-p-\xi_C(1-2p))\end{aligned}$$

Thus,

$$\overline{p}(\mu, p) = \mu_S(\beta p_S + (1-\beta)p) + \mu_C(1-p-\xi_C(1-2p)) \tag{5}$$

From this we can derive the expression for p as a function of p_S

$$\begin{aligned}\bar{p} &= \mu_S\beta p_S + \mu_S(1-\beta)p + \mu_C - \mu_C p - \mu_C\xi_C + 2\mu_C p\xi_C \\ \bar{p} &= \mu_S\beta p_S + \mu_C - \mu_C\xi_C + p[\mu_S(1-\beta) - \mu_C(1-2\xi_C)] \\ &= \mu_S\beta p_S + \mu_C - \mu_C\xi_C + p[\mu_S(1-\beta) - \mu_C(1-2\xi_C)]\end{aligned}$$

Thus p is defined as

$$p = \frac{\bar{p} - \mu_S\beta p_S - \mu_C(1-\xi_C)}{\mu_S(1-\beta) - \mu_C(1-2\xi_C)} \tag{6}$$

Finally, we can define the parameter $\hat{p}$ that is defined as the expectation of p conditioning on $\bar{p}$

$$\hat{p}\& = \frac{\bar{p} - [\mu_S\beta\mathbb{E}[p_S|\bar{p}] + \mu_C(1-\xi_C)]}{\mu_S(1-\beta) + \mu_C(2\xi_C - 1)} \tag{7}$$

Thus, $\bar{p}$ is defined as

$$\hat{p} = \mathbb{E}[p|\alpha_1 p + \alpha_2 p_S = \bar{p}] \tag{8}$$

$$\begin{aligned}\hat{p} &= \frac{\beta^2\mu_S^2\mu_X\sigma_Y^2 + \sigma_X^2(1-\bar{p}-\xi_C+\mu_S(-1+\beta\mu_Y+\xi_C))(1-2\xi_C+\mu_S(-2+\beta+2\xi_C))}{\beta^2\mu_S^2\sigma_Y^2 + \sigma_X^2(1-2\xi_C+\mu_S(-2+\beta+2\xi_C))^2} \\ &= \frac{\beta^2\mu_S^2\mu_X\sigma_Y^2 + \sigma_X^2(\bar{p}-[\mu_S\beta\mu_y+(1-\mu_S)(1-\xi_C)])(1-2\xi_C+\mu_S(-2+\beta+2\xi_C))}{\beta^2\mu_S^2\sigma_Y^2 + \sigma_X^2(1-2\xi_C+\mu_S(-2+\beta+2\xi_C))^2}\end{aligned}$$

2) It is worth noting that the NWF can be expressed as the sum of the PWF is a

biased term due to the elections. In fact,

$$\begin{aligned}
NWF &= -\mathbb{E}\left[(p-\bar{p})^2\right] = -\mathbb{E}\left[(p-\hat{p}+\hat{p}-\bar{p})^2\right] \\
&= -\mathbb{E}\left[(p-\hat{p})^2 + 2(p-\hat{p})(\hat{p}-\bar{p}) + (\hat{p}-\bar{p})^2\right] \\
&= -\left[\mathbb{E}\left[(p-\hat{p})^2\right] + 2\mathbb{E}\left[(p-\hat{p})(\hat{p}-\bar{p})\right] + \mathbb{E}\left[(\hat{p}-\bar{p})^2\right]\right] \\
&= -\left[\mathbb{E}\left[(p-\hat{p})^2\right] + 2(\hat{p}-\bar{p})\underbrace{\mathbb{E}\left[(p-\hat{p})\right]}_{0} + \mathbb{E}\left[(\hat{p}-\bar{p})^2\right]\right] \\
&= -\left[\underbrace{\mathbb{E}\left[(p-\hat{p})^2\right]}_{\text{Precision of elections}} + \underbrace{\mathbb{E}\left[(\hat{p}-\bar{p})^2\right]}_{\text{Bias of elections}}\right]
\end{aligned}$$

Thus, it can be rewritten as

$$NWF = PWF + Bias$$

At this stage, we define the two welfare functions given the distributions of the parameters.

3)What the analysis aims showing is the evolution of the welfare functions over time and the main differences between the evolution of the PWF and the NWF. In particular, to study the evolution, we take the first derivative of the two functions with respect to μ_S. It is necessary and sufficient to show the sign of this derivative in order to have an all-rounded understanding of the evolution of the two functions. In fact, μ_S as defined above depends negatively on t and λ_1. Hence, once we define the relation between the functions and μ_S, we immediately get to know the relation between the functions and the time/lambda. Thus, let us start by showing the behavior of the PWF.

$$\frac{\partial PWF}{\partial \mu_S} = \frac{2\beta^2\mu_S\sigma^2\left(-1+2\xi_C\right)\left[1-2\xi_C+\mu_S\left(-2+\beta+2\xi_C\right)\right]}{\left\{2\mu_S^2\left[\beta^2+2\beta\left(-1+\xi_C\right)+2\left(-1+\xi_C\right)^2\right]+\left(1-2\xi_C\right)^2-2\mu_S\left(-1+2\xi_C\right)\left(-2+\beta+2\xi_C\right)\right\}^2} \propto 1-2\xi_C+\mu_S\left(\beta-2\left(1-\xi_C\right)\right)$$

Since almost everything is bigger than or equal to 0, if we want to study the sign of the above formula, then we have to analyze the sign of the following term.

$$1 - 2\xi_C + \mu_S\left(-2 + \beta + 2\xi_C\right) < 0$$
$$0 < \mu_S < \underbrace{\frac{2\xi_C - 1}{-2 + \beta + 2\xi_C}}_{\geq 1?}$$

Proposition 3 *W_P is increasing in t and λ if*

$$\frac{2\xi_C - 1}{-2 + \beta + 2\xi_C} > 1 \iff 1 > \beta$$

Let us study the right-hand side of the inequality.

$$2\xi_C - 1 \geq -2 + \beta + 2\xi_C$$
$$\beta \leq 1$$

Therefore, we can conclude that PWF decreases in μ_S each time t, because

$$0 < \mu_S < \frac{2\xi_C - 1}{-2 + \beta + 2\xi_C}, \quad \forall \mu_S \in [0, 1]$$

In other words, the PWF is an increasing function of both t and λ_1.

A.2 Proof of Proposition 1

Proof. Let $\eta = \eta_N$ be the share of citizens who have not become critical thinkers yet. Using the chain rule, we obtain both welfare functions.

$$\frac{\mathrm{d}W}{\mathrm{d}\lambda} = \frac{\mathrm{d}W}{\mathrm{d}\eta} \cdot \underbrace{\frac{\mathrm{d}\eta}{\mathrm{d}\lambda}}_{<0} \implies \frac{\mathrm{d}W}{\mathrm{d}\lambda} \propto -\frac{\mathrm{d}W}{\mathrm{d}\eta}$$

welfare moves in contrary to how it moves in η, which in turns decreases in λ (and t). First, for positive welfare, $\frac{\mathrm{d}W}{\mathrm{d}\eta}$ is always negative for the following computations.

$$\frac{\partial PWF}{\partial \mu_S} =$$

$$\frac{2\beta^2\mu_S\sigma^2\left(-1+2\xi_C\right)\left[1-2\xi_C+\mu_S\left(-2+\beta+2\xi_C\right)\right]}{\left\{2\mu_S^2\left[\beta^2+2\beta\left(-1+\xi_C\right)+2\left(-1+\xi_C\right)^2\right]+\left(1-2\xi_C\right)^2-2\mu_S\left(-1+2\xi_C\right)\left(-2+\beta+2\xi_C\right)\right\}^2}$$

$$\propto 1-2\xi_C+\mu_S\left(\beta-2\left(1-\xi_C\right)\right)$$

Since almost everything is bigger or equal to 0, if we want to study the sign of the above formula, then we have to analyze the sign of the following term

$$1-2\xi_C+\mu_S\left(-2+\beta+2\xi_C\right)<0$$

$$0<\mu_S<\underbrace{\frac{2\xi_C-1}{-2+\beta+2\xi_C}}_{\geq 1?}$$

Let's study the right-hand side of the inequality

$$2\xi_C-1\geq -2+\beta+2\xi_C$$

$$\beta\leq 1$$

Therefore, we can conclude that PWF is decreasing in μ_S for each time t, because

$$0<\mu_S<\frac{2\xi_C-1}{-2+\beta+2\xi_C},\quad \forall \mu_S\in[0,1]$$

Furthermore, $\frac{\mathrm{d}W}{\mathrm{d}\eta}$ has a non-trivial solution. That is, by studying the sign of the derivative of welfare elections with respect to μ_S we obtain

$$\frac{\partial NWF}{\partial \mu_S}=-\left[4\beta^2\mu_S\sigma^2+4\beta\left(-1+\mu_S\right)\sigma^2\left(-1+\xi_C\right)+4\beta\mu_S\sigma^2\left(-1+\xi_C\right)+2\left(-1+\mu_S\right)\left[\left(1-2\mu\right)^2+4\sigma^2\right]\left(-1+\xi_C\right)^2\right]$$

The sign of the term in brackets is

$$4\beta^2\mu_S\sigma^2+4\beta\sigma^2\left(-1+\xi_C\right)\left(-1+2\mu_S\right)+2\left(-1+\mu_S\right)\left[\left(1-2\mu\right)^2+4\sigma^2\right]\left(-1+\xi_C\right)^2>0$$

$$\mu_S\left[4\beta^2\sigma^2+8\beta\sigma^2\left(-1+\xi_C\right)+2\left(-1+\xi_C\right)^2\left[\left(1-2\mu\right)^2+4\sigma^2\right]\right]>4\beta\sigma^2\left(-1+\xi_C\right)+2\left(-1+\xi_C\right)^2\left[\left(1-2\mu\right)^2+4\sigma^2\right]$$

$$\mu_S>\underbrace{\frac{4\beta\sigma^2\left(-1+\xi_C\right)+2\left(-1+\xi_C\right)^2\left[\left(1-2\mu\right)^2+4\sigma^2\right]}{4\beta^2\sigma^2+8\beta\sigma^2\left(-1+\xi_C\right)+2\left(-1+\xi_C\right)^2\left[\left(1-2\mu\right)^2+4\sigma^2\right]}}_{\text{Threshold}<1?}$$

Saying that the threshold is less than one also means that $\frac{\partial NWF}{\partial \mu_S}<0 \iff \mu_S>$

Threshold. We want to study this threshold. The conditions given in the text correspond to this threshold being below 0 and above 1, respectively.

$$\begin{cases} COND1 \rightarrow \text{Threshold} < 0 & \text{Always decreases in time} \\ COND2 \rightarrow \text{Threshold} > 1 & \text{Always increase in time} \\ COND3 \rightarrow \text{Threshold} \in (0,1) & \text{Increases first, decreases later} \end{cases}$$

COND1 occurs according to the following expression

$$\beta > \frac{(1-\xi_C)\left((1-2\mu)^2+4\sigma^2\right)}{2\sigma^2}$$

Then, the welfare is always decreasing. For COND2 to occur, the numerator of the threshold must be higher than the denominator. Hence, since only two terms differ between numerator and denominator, the following must be true.

$$\begin{aligned} 4\beta\sigma^2(-1+\xi_C) &> 4\beta^2\sigma^2 + 8\beta\sigma^2(-1+\xi_C) \\ 4\beta\sigma^2(-1+\xi_C) &> 4\beta\sigma^2(\beta+2\xi_C-2) \\ \beta &< 1-\xi_C \end{aligned}$$

The welfare is then always increasing. Finally, COND3 can be intuitively discussed. Since μ_S is monotonically decreasing in time, there must be continuity of t^{max} such that

$$\begin{cases} \frac{\partial W^N}{\partial \mu_S} < 0 & \text{for } t < t^{max} \\ \frac{\partial W^N}{\partial \mu_S} < 0 & \text{for } t > t^{max} \end{cases}$$

Therefore, t^{max} is a maximum interior of W^N when threshold $\in (0,1)$ ■

A.3 Proof of Proposition 2

Proof. Firstly, define the parameters associated with $\bar{p}$

$$\bar{p} = \mu_T p + \mu_S(\beta p_S + (1-\beta)p) + \mu_C[1-p+\xi_C(2p-1)]$$

where

$$\alpha_0 = \mu_C (1 - \xi_C)$$
$$\alpha_1 = 1 - \beta\mu_S - 2\mu_C (1 - \xi_C)$$
$$\alpha_2 = \beta\mu_S$$

Then $\hat{p}$ is given by

$$\hat{p} = \frac{\frac{\bar{p} - [\alpha_0 + \alpha_2 \mu_y]}{\alpha_1} \alpha_1^2 \sigma_x^2 + \alpha_2^2 \sigma_y^2 \mu_x}{\alpha_1^2 \sigma_x^2 + \alpha_2^2 \sigma_y^2}$$

where

$$\gamma_0(t, \lambda) = \frac{\alpha_2^2 \sigma_y^2 \mu_x - \alpha_2 \alpha_1 \sigma_x^2 \mu_y}{\alpha_1^2 \sigma_x^2 + \alpha_2^2 \sigma_y^2}$$
$$\gamma_1 (t, \lambda) = \frac{\alpha_1^2 \sigma_x^2}{\alpha_1^2 \sigma_x^2 + \alpha_2^2 \sigma_y^2}$$
$$\gamma_2(t, \lambda) = \frac{\alpha_2 \alpha_1 \sigma_x^2}{\alpha_1^2 \sigma_x^2 + \alpha_2^2 \sigma_y^2}$$

The bias is zero if and only if the following system has a solution.

$$\begin{cases} \alpha_1 = \gamma_1 \\ \alpha_2 = \gamma_2 \end{cases}$$

that is

$$\begin{cases} \alpha_1 = \frac{\alpha_1^2 \sigma_x^2}{\alpha_1^2 \sigma_x^2 + \alpha_2^2 \sigma_y^2} \\ \alpha_2 = \frac{\alpha_2 \alpha_1 \sigma_x^2}{\alpha_1^2 \sigma_x^2 + \alpha_2^2 \sigma_y^2} \end{cases}$$

It is immediate to check that $\gamma_1 (t, \lambda) = \alpha_1 (t, \lambda) \iff \gamma_2 (t, \lambda) = \alpha_2 (t, \lambda)$, so, we actually have a single equation, and we need to claim that it exists a time such that

$$\begin{aligned}
\frac{\alpha_1^2\sigma_x^2}{\alpha_1^2\sigma_x^2+\alpha_2^2\sigma_y^2}=\alpha_1 &\iff \frac{\left(1-\beta\mu_S-2\mu_C\left(1-\xi_C\right)\right)^2\sigma_x^2}{\left(1-\beta\mu_S-2\mu_C\left(1-\xi_C\right)\right)^2\sigma_x^2+\left(\beta\mu_S\right)^2\sigma_y^2}=\left(1-\beta\mu_S-2\mu_C\left(1-\xi_C\right)\right)\\
&\iff \left(1-\beta\mu_S-2\mu_C\left(1-\xi_C\right)\right)\sigma_x^2=\left(1-\beta\mu_S-2\mu_C\left(1-\xi_C\right)\right)^2\sigma_x^2+\left(\beta\mu_S\right)^2\sigma_y^2\\
&\iff \left(1-\beta\mu_S-2\mu_C\left(1-\xi_C\right)\right)\sigma_x^2\left(\beta\mu_S+2\mu_C\left(1-\xi_C\right)\right)=\left(\beta\mu_S\right)^2\sigma_y^2\\
&\iff \frac{\left(1-\beta\mu_S-2\mu_C\left(1-\xi_C\right)\right)\left(\beta\mu_S+2\mu_C\left(1-\xi_C\right)\right)}{\left(\beta\mu_S\right)^2}=\frac{\sigma_y^2}{\sigma_x^2}\\
&\iff \frac{\left(1-\beta\mu_S-2\left(1-\mu_S\right)\left(1-\xi_C\right)\right)\left(\beta\mu_S+2\left(1-\mu_S\right)\left(1-\xi_C\right)\right)}{\left(\beta\mu_S\right)^2}
\end{aligned}$$

When $\mu_S=0$ there cannot be the zero-bias time, because as $t\to\infty$ this explodes (? can we show this is always increasing in μ_S) because there is always bias in the limits. On the other hand, there could be a zero-bias time that coincides with $t^\star=0$. Indeed, when $\mu_S=1$

$$\frac{1-\beta}{\beta}=\frac{\sigma_y^2}{\sigma_x^2}$$

An even more special case is when $\xi_C=1$. Indeed,

$$\begin{aligned}
\frac{\left(1-\beta\mu_S\right)^2\sigma_x^2}{\left(1-\beta\mu_S\right)^2\sigma_x^2+\left(\beta\mu_S\right)^2\sigma_y^2}&=\left(1-\beta\mu_S\right)\\
\left(1-\beta\mu_S\right)\sigma_x^2&=\left(1-\beta\mu_S\right)^2\sigma_x^2+\left(\beta\mu_S\right)^2\sigma_y^2\\
\left(\frac{1-\beta\mu_S}{\beta\mu_S}\right)&=\frac{\sigma_y^2}{\sigma_x^2}
\end{aligned}$$

Substituting the expression of $\mu_S$as a function of t and λ

$$\frac{\sigma_x^2}{\beta\left(\sigma_x^2+\sigma_y^2\right)}=e^{-t\lambda_1}$$

that becomes

$$t^\star=-\frac{1}{\lambda_1}\log\left(\frac{\sigma_x^2}{\beta\left(\sigma_x^2+\sigma_y^2\right)}\right)$$

Where the argument of the log must be smaller than 1

$$\frac{\sigma_x^2}{\beta\left(\sigma_x^2+\sigma_y^2\right)}<1$$

that is

$$\frac{(1-\beta)}{\beta} < \frac{\sigma_y^2}{\sigma_x^2}$$

■

B Experimental Details

B.1 Data Collection

Preventing duplicates. Submissions to studies on Prolific are guaranteed to be unique by the firm.[21] Our system is set up so that each participant can submit only one per study on Prolific. That is, each participant will be listed in our dashboard only once and can only be paid once. On our side, we also prevent participants from taking the experiment several times in two steps. First, we enable the functionality "Prevent Ballot Box Stuffing," which prevents multiple entries from the same user by tracking IP addresses, setting browser cookies, and requiring unique account verifications. This ensures that each participant can only submit their response once. Second, we check for unique participant IDs in the dataset and delete duplicate submissions if we find any.

High vs low-quality submissions. Participants joining the Prolific pool receive a rate based on the quality of their engagement with the studies. If they are rejected from a study, then they will receive a negative score. If they receive too many negative scores, then Prolific removes them from their pool of potential participants for different studies.[22] Based on this long-term contract, participants are incentivized to pay attention and follow the expectations of each study. Hence, good research behavior has emerged on Prolific, according to which participants themselves can vol voluntarily withdraw their submissions if they feel they made a mistake, such as rushing too much, letting the survey open for a long period without engaging with it, and so on.* According to these standards, we kept submission rejections as low as possible, following the standard in online experimental economics. Participants who fail at least one fair attention check are rejected and not paid. Following Prolific standards, participants who are statistical outliers (3 standard deviations below the mean) are excluded from the good complete data set.

[21] See Prolific unique submission guarantee policy here.
[22] See Prolific pool removal Policy here.

Payments and communication. We make sure to review participants' submissions within 24-48 hours after they have completed the study. This means that within this time frame, if we accept their submission, they will receive their fixed and bonus payment. Otherwise, we reject their submissions and send them a personalized e-mail detailing the reason for the rejection, leaving participants the opportunity to contact us afterward if they firmly believe the decision to be unfair (motivating their perspective). Participants can also contact us at any time if they encounter problems with our study or have questions about it.

B.2 Cheating Behavior Monitoring Algorithms

The first algorithm tracks the number of times participants open a new tab on their computer during the essay exercise and how much time they spend on our essay writing web page. We gathered only the following information: 'participant i has opened a new tab during the essay, n number of times, for some time t.

The second algorithm checks whether participants copy and paste external information by comparing the number of written characters and the number of keyboard clicks. If the number of keyboard clicks is strictly less than the number of written characters, this implies that the participants have copied external information.

This second algorithm cannot distinguish between the original external information and plagiarism. In the situation in which some participants had already written about a topic or a relevant topic and saved it on their computer before coming to the experiment, we also used a feature in the AI software to check for plagiarism after the participants had finished the experiment.

B.3 Generating Storytelling Contexts with LLMs

Description of the storytelling contexts. First, in the *Newspaper* treatment, participants are exposed to two news articles: one for and one against the issue. Each article is presented for 60 seconds. This context embodies a structured and formal writing style, minimalist visual design, high source credibility from an institutional perspective, and content delivery that presents a continuous, long-form story within a single screen frame.

Second, in the *Facebook* treatment, participants are exposed to six Facebook posts: two for and two against the issue, as well as two irrelevant posts. Each post is displayed for 20 seconds. This context features a moderately formal writing style with variable length, a colorful and interactive visual design, source credibility from personal and diverse user-generated content, and content delivery that presents indi-

vidual stories sequentially, one per screen frame.

Third, in the *Twitter* treatment, participants are exposed to twenty-four tweets: ten for digital privacy, ten against digital privacy, and four ads. Each tweet has an average length of 40 characters and is displayed for five seconds, aligning with the average reading speed in the US population. This context is characterized by a concise and informal writing style, minimalist visual design, varied source credibility from multiple anonymous individuals, and content delivery that rapidly presents individual stories in quick succession, one per screen frame.

Fourth, an additional storytelling context was designed: *Biased Twitter*. In this treatment, participants are exposed to thirteen tweets: ten for and three ads or ten against and three ads. Each tweet is displayed for five seconds. This context maintains the concise and informal writing style of tweets, minimalist visual design, and varied source credibility from multiple anonymous individuals. The content delivery rapidly presents individual stories in quick succession, one per screen frame. The *Biased Twitter* treatment was introduced to specifically analyze the effects of exposure to a one-sided (biased) storytelling context, contrasting with the two-sided storytelling context in the *Twitter* treatment. By comparing the two, we aim to understand which role exposure to a predominantly one-sided storytelling context can play in prompting or not critical thinking in users.

Treatment	Description	Content	Individual Exposure Time
Newspaper	Participants are exposed to two news articles: one for and one against the issue.	2 news articles	180 seconds
Facebook	Participants are exposed to six Facebook posts: two for, two against the issue, and two irrelevant posts.	6 Facebook posts	60 seconds per post
Twitter	Participants are exposed to twenty-four tweets: ten for digital privacy, ten against digital privacy, and four ads. Each tweet has an average length of 40 characters.	24 tweets	15 seconds per tweet
Biased Twitter	Participants are exposed to thirteen tweets: ten for and three ads or ten against and three ads.	12 tweets	15 seconds per tweet

Table 4: Summary of Treatment Details

How we used LLMs to generate the storytelling contexts. First, once we extracted all the facts used in our experiment from the newspaper articles, we used a combination of two LLMs, Quillbot and Copy.AI, to aid in paraphrasing a crude and short writing style, as on Twitter and a medium writing style, as on Facebook. Second, we relied on Zeeob, a social media format simulator, to generate UX design elements such as fonts, line spacing, color, and other elements specific to both storytelling contexts, *Twitter* and *Facebook*. Third, for the credibility sources, based on the UX design, the participants are immediately exposed to Twitter and Facebook branding. For the *Newspaper* context, participants were told that the article was from a leading newspaper. In addition to branding, another source of credibility, particularly in the case of social media, is how many different individuals share similar stories (facts going in the same direction, for or against). To generate these individuals, we used StyleGAN2, an image generation machine learning model, to generate deep fakes, which were then blurred to keep participants' attention focused on how different individuals were sharing stories and not sharing them with *who* (in terms of gender or ethnic stereotypes) was sharing them.[23]

[23] Access the respective AI tools we used here: Quillbot: https://quillbot.com/; CopyAI: https://www.copy.ai/; Zeeob: https://zeoob.com/; StyleGAN2: https://github.com/NVlabs/stylegan2.

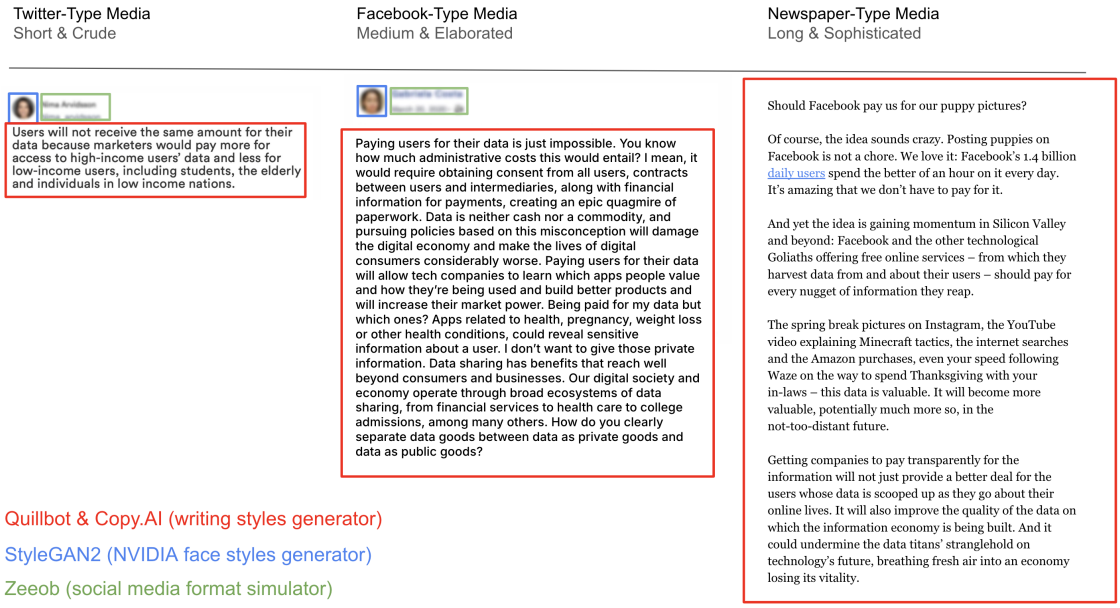


Figure 3: Examples of Twitter, Facebook, and Newspaper treatments

B.4 Grading Essays with LLMs

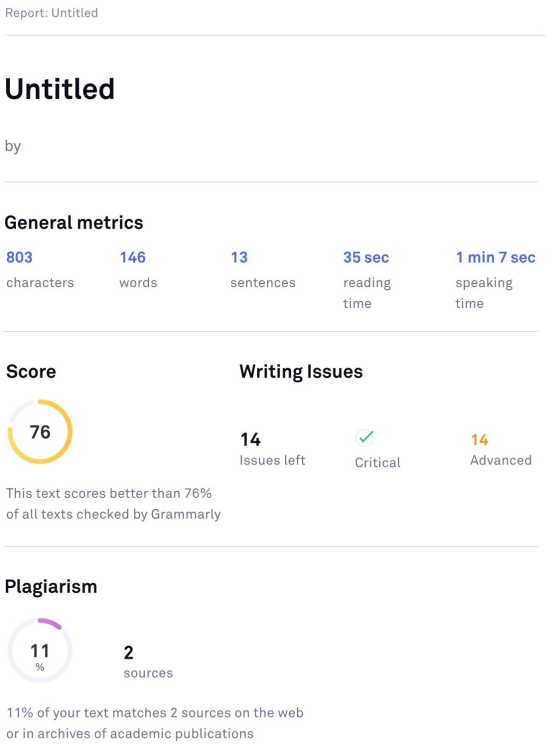


Figure 4: Example of Grammarly-generated grade report

B.5 Detailed Elicitations

B.5.1 Political Preferences

Ex-ante and ex-post political preferences. Before the treatment, we prompt participants on different political issues (i.e., without a baseline): guns, crime, climate, welfare, and digital privacy issues. We use the standard congressional metrics, including digital issues. We elicit more than only digital preferences to ensure that participants do not guess at this stage which preferences we focus on in the remainder of the experiment (treatment and critical thinking essay) to minimize their social desirability bias. After the treatment on digital privacy, we survey participants again to elicit their preferences about digital privacy. We use the following scale.

1. On the issue of gun regulation, do you support or oppose the following proposal?
2. On the issue of environmental policies, do you support or oppose the following proposal?
3. On the issue of crime policies, do you support or oppose the following proposal?
4. On the issue of digital policies, do you support or oppose each of the following proposals?

B.5.2 Digital Knowledge Test

See the participants' experimental instructions.

B.5.3 Issue Familiarity

1. In the remainder of the experiment, we will focus on the following political issue [text]. Please, again, state your preference.
2. Have you thought deeply about this issue before participating in this study? [Yes/No]

B.5.4 Listing Reasons

If yes to the previous question, then participants see this question:

You answered "Yes" to the previous question. You will be asked now to provide, at most, two reasons that justify your position and two reasons that justify the opposite position. If you do not know any reasons, please select "I am unable to list

any logical reason at the moment". You do not need to agree with these reasons; they need to be a logical justification for or against your position. Your payment WILL NOT depend on your answer to this question. However, your honest answer is of paramount importance for the success of this study.

1. Reasons which justify your position
    - Reason 1: [write text here]
    - Reason 2: [write text here]
    - I am unable to list any logical reason at the moment
2. Reasons which oppose your position
    - Reason 1: [write text here]
    - Reason 2: [write text here]
    - I am unable to list any logical reason at the moment

B.5.5 Internal Uncertainty

How certain are you of your preference regarding the digital privacy issue? By "Certain", we mean that you feel confident enough to vote for your political preference if asked to you in a real-life political committee. Select among the following options:

- Completely Uncertain
- Rather Uncertainty
- Rather Certain
- Completely Certain

B.5.6 Need for Cognition

For each sentence below, select how uncharacteristic or characteristic this is for you.

Note: Despite its wide use, to our knowledge, no average for the US population is available. Originally a 34-question version, the authors developed an 18-question version for efficiency. We use here a validated 6-question version more adapted for online artefactual experiments.

B.5.7 Cognitive Flexibility

Note: The author provide the average of 55.

B.6 Critical Thinking Classification

B.6.1 Critical Thinking Classification Results

Table 5 shows the classification results of individuals as *N*aive and Critical Thinking.

Treatment	$N_0 \to N_1$	$N_0 \to C_1$	$C_0 \to C_1$
NEWSPAPER	107	46	17
TWITTER	130	42	18
BIASED TWITTER	111	51	22
FACEBOOK	109	59	13
N	368	258	99

Table 5: Pre- and Post-Treatment Classification

B.6.2 Distribution of Cognitive Styles

Table 3 presents differences in the mental push to critical thinking through storytelling among participants with high cognition needs. In this section, we provide the experimental results among participants with a low need for cognition, high cognitive flexibility, and low cognitive flexibility.

Treatment	*Newspaper*	*Twitter*	BIASED*Twitter*	*Facebook*
Newspaper	·	·	·	·
Twitter	0.888 (0.067)	·	·	·
Biased Twitter	0.736 (0.068)	−0.136 (0.065)	·	·
Facebook	0.499 (0.068)	−0.398 (0.064)	−0.253 (0.066)	·
N	99	113	103	109

Standard errors in parentheses
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 6: *Z-Scores* for a Low Need for Cognition

C Robustness Checks

C.1 Sensitivity Analysis

This appendix provides detailed statistical results supporting the main findings discussed in the paper. The tables below present the z-scores and p-values for different threshold sensitivities and levels of cognition, focusing on *Facebook*'s effectiveness

Treatment	*Newspaper*	*Twitter*	*Biased Twitter*	*Facebook*
Newspaper	·	·	·	·
Twitter	0.723 (0.078)	·	·	·
Biased Twitter	−0.231 (0.087)	−0.963 (0.080)	·	·
Facebook	−0.979 (0.089)	−1.751* (0.082)	−0.745 (0.090)	·
N	59	68	56	57

Standard errors in parentheses
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 7: *Z-Scores* for High Cognitive Flexibility

Treatment	*Newspaper*	*Twitter*	BIASED*Twitter*	*Facebook*
Newspaper	·	·	·	·
Twitter	0.850 (0.063)	·	·	·
Biased Twitter	−0.143 (0.064)	−1.015 (0.062)	·	·
Facebook	−0.497 (0.064)	−1.387 (0.061)	−0.361 (0.063)	·
N	111	122	128	124

Standard errors in parentheses
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 8: *Z-Scores* for Low Cognitive Flexibility

in fostering critical thinking compared to other platforms such as *Twitter* and Newspapers.

Overall, the data supports our conclusion that *Facebook* is an effective platform for fostering critical thinking across various thresholds, especially when participants have higher baseline knowledge. These results highlight the importance of minimal knowledge as a necessary condition for the cognitive engagement process, aligning with the literature that emphasizes the need for foundational knowledge in the development of critical thinking. This suggests a targeted approach in educational and policy-making strategies to maximize the benefits of media-driven cognitive engagement.

Table 9 shows the comparative effectiveness of *Facebook*, *Twitter*, and *Newspaper* on different knowledge thresholds for participants with a high need for cognition. The data indicates that *Facebook* consistently outperforms the other platforms, especially at higher thresholds (9 and 10 correct answers), where it shows significant advantages over both *Twitter* and Newspapers. This demonstrates Facebook's effectiveness in promoting critical thinking, particularly when participants have substantial baseline knowledge.

Threshold	*Newspaper*	*Twitter*	*Biased Twitter*
Facebook (10, both)	-2.438^{**}	-2.968^{***}	-1.228
	(0.078)	(0.074)	(0.078)
Facebook (9, both)	-2.451^{**}	-3.223^{***}	-1.256
	(0.080)	(0.075)	(0.081)
Facebook (5, both)	-1.601	-2.674^{***}	-0.821
	(0.087)	(0.080)	(0.085)
Facebook (8, 3)	-2.229^{**}	-3.051^{***}	-0.970
	(0.081)	(0.076)	(0.082)
Facebook (8, 1)	-2.309^{**}	-3.560^{***}	-0.930
	(0.083)	(0.077)	(0.085)

Table 9: *Z-Score* for Threshold Sensitivity and High Need for Cognition

Table 10 provides results for the broader participant base, illustrating how *Facebook* compares to *Twitter* and *Newspaper* across all levels of cognition. The results reinforce the findings that *Facebook* remains the most effective platform for encouraging critical thinking, even when varying levels of baseline knowledge are taken into account. In particular, the advantage of *Facebook* is less pronounced at the lowest threshold (5 correct answers), underscoring the need for minimal baseline knowledge as a prerequisite for effective critical thinking engagement.

Threshold	*Newspaper*	*Twitter*	Biased *Twitter*
Facebook (10, both)	−1.066	−2.122**	−1.013
	(0.050)	(0.047)	(0.049)
Facebook (9, both)	−1.106	−2.368**	−0.996
	(0.052)	(0.048)	(0.051)
Facebook (5, both)	−0.458	−1.811*	−0.567
	(0.054)	(0.051)	(0.053)
Facebook (8, 3)	−0.882	−2.200**	−0.855
	(0.052)	(0.049)	(0.051)
Facebook (8, 1)	−1.096	−2.467**	−0.713
	(0.054)	(0.050)	(0.054)

Table 10: *Z-Score* for Threshold Sensitivity for all Levels of Cognition

C.2 News Consumption Habits

C.2.1 Table of Daily Effects For General and High Need Cognition

Habit	*Newspaper*	*Twitter*	*Facebook*
Newspaper	·	0.153	−0.222*
		(0.115)	(0.128)
Twitter	·	·	−0.375***
			(0.120)
Facebook	·	·	·
N	128	142	128

Standard errors in parentheses
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 11: *Z-Scores* for News Consumption Habits: Daily

Habit	*Newspaper*	*Twitter*	*Facebook*
Newspaper	·	0.155	−0.371
		(0.176)	(0.201) *
Twitter	·	·	−0.526
			(0.189) ***
Facebook	·	·	·
N	53	57	53

Standard errors in parentheses
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 12: *Z-Scores* for High Need for Cognition Participants with Daily Habits

C.2.2 Different Timeline of News Consumption Habits

No Effect for Weekly Habits. The analysis of weekly news consumption habits reveals no statistically significant differences in effectiveness between the platforms examined. Specifically, when comparing *Newspaper* and *Facebook*, the Z-score difference was 0.420 ($p = 0.146$), indicating no significant advantage for *Newspaper* over *Facebook* in providing political information. Similarly, the comparison between *Facebook* and *Twitter* yielded a Z-score difference of 0.173 ($p = 0.504$), highlighting that *Facebook* does not significantly outperform *Twitter* for weekly users. These findings suggest that weekly news consumers do not exhibit a strong preference for one platform over another in acquiring political information.

No Effects for Monthly Habits. The analysis of monthly news consumption habits reveals no statistically significant differences in effectiveness among Newspaper, Twitter, Biased Twitter, and *Facebook* for delivering political content. Specifically, the comparison between *Newspaper* and *Facebook* showed a Z-score difference of 0.000 ($p = 1.000$), indicating complete parity in their effectiveness. Similarly, the *Twitter* and *Facebook* comparison yielded a Z-score difference of 0.171 ($p = 0.644$), demonstrating that neither platform significantly outperforms the other. Overall, these findings suggest that for consumers who engage with news on a monthly basis, no specific platform offers a distinct advantage in political information dissemination.

C.2.3 Social Media Platform Habits

Facebook habits. The analysis of participants who marked "Facebook" hows no significant differences in Z-scores between Newspaper, Twitter, and Facebook. Specifically, the comparison between **Newspaper** and **Twitter** reveals a Z-score difference of 0.159 ($p = 0.364$), indicating parity in their effectiveness. Similarly, the comparison between **Newspaper** and **Facebook** yields a Z-score difference of 0.109 ($p = 0.549$), and **Twitter** and **Facebook** show a Z-score difference of -0.051 ($p = 0.759$). These findings suggest that for this group, no platform significantly outperforms the others in delivering political information.

Twitter habits. The analysis of participants who marked "Twitter" in habit 2_1 shows no significant differences in Z-scores between Newspaper, Twitter, and Facebook. Specifically, the comparison between **Newspaper** and **Twitter** reveals a Z-score difference of 0.004 ($p = 0.979$), indicating parity in their effectiveness. Similarly, the comparison between **Newspaper** and **Facebook** yields a Z-score difference of -0.154 ($p = 0.365$), and **Twitter** and **Facebook** show a Z-score difference of

-0.158 ($p = 0.363$). These findings suggest that for this group, no platform significantly outperforms the others in delivering political information.

Other and none of the above platforms habits. The analysis of participants who marked "Others" shows no significant differences in Z-scores between Newspaper, Twitter, and Facebook. Specifically, the comparison between **Newspaper** and **Twitter** reveals a Z-score difference of -0.006 ($p = 0.972$), indicating parity in their effectiveness. Similarly, the comparison between **Newspaper** and **Facebook** yields a Z-score difference of -0.252 ($p = 0.143$), and **Twitter** and **Facebook** show a Z-score difference of -0.247 ($p = 0.149$). These findings suggest that for this group, no platform significantly outperforms the others in delivering political information.

The analysis of participants who marked "None of the above" shows no significant differences in Z-scores between Newspaper, Twitter, and Facebook. Specifically, the comparison between **Newspaper** and **Twitter** reveals a Z-score difference of 0.276 ($p = 0.188$), indicating parity in their effectiveness. Similarly, the comparison between **Newspaper** and **Facebook** yields a Z-score difference of 0.022 ($p = 0.920$), and **Twitter** and **Facebook** show a Z-score difference of -0.253 ($p = 0.203$). These findings suggest that for this group, no platform significantly outperforms the others in delivering political information.

C.3 Human Expert Feedback

C.3.1 Replacing Human Expertise by Essay Length

This appendix provides detailed statistical analyses that support the findings discussed in the main text. The tables below include results on the effectiveness of different media platforms in fostering critical thinking, the accuracy of psychologists in identifying AI-generated essays, and the hedging effects related to AI essay identification.

Length-Based Metrics to Detect Critical Thinking are misleading. Table 13 presents z-scores that compare the effectiveness of various media platforms in promoting critical thinking, measured by the proportion of reasoning styles influenced by different treatments. The results show that *Facebook*'s impact on promoting critical thinking is less significant than *Twitter* and *Newspaper* treatments. For example, *Facebook* showed a weaker influence compared to *Twitter* ($z = 1.843^{*}; p < 0.1$), emphasizing the need for human evaluation over simple length-based metrics.

Treatment	*Newspaper*	*Twitter*	*Biased Twitter*	*Facebook*
Newspaper	·	·	·	·
Twitter	−2.051**	·	·	·
	(0.089)			
Biased Twitter	−1.466	0.635	·	·
	(0.085)	(0.091)		
Facebook	−0.171	1.843*	1.262	·
	(0.086)	(0.091)	(0.088)	
N	71	77	81	72

Standard errors in parentheses
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 13: Z-Scores for Length as Ex-Post and High Need for Cognition

Accuracy Rate by Psychologists to Detect AI-generated Essays and Hedging Effects. Table 14 below summarizes the accuracy rates of psychologists in distinguishing between AI-generated and human-written essays. The table highlights the variability in accuracy, with rates ranging from 59% to 92%. The weak positive correlation ($r = 0.25$) between the number of AI essays evaluated and the accuracy of the identification suggests that experience may slightly improve detection capabilities, underscoring the importance of human expertise in evaluating cognitive skills.

ID	Correct Answers	AI Essays	Total Pass	Total Essays	Accuracy Rate	Pass Rate
1	90	22	34	122	0.74	0.28
2	70	22	48	119	0.59	0.40
3	98	22	38	122	0.80	0.31
4	56	11	17	61	0.92	0.28
5	196	44	68	238	0.82	0.29
6	89	22	34	112	0.79	0.30
7	46	11	11	61	0.75	0.18
8	102	22	38	122	0.84	0.31
9	92	22	27	112	0.82	0.24
10	45	11	26	61	0.74	0.43
11	94	22	19	120	0.78	0.16
12	94	22	71	122	0.77	0.58
13	43	11	28	61	0.70	0.46
15	41	11	9	51	0.80	0.18
NA	4	0	1	31	0.67	0.03

Table 14: Accuracy Rate by Psychologists

Table 15 examines the hedging effects related to the accuracy of identifying AI-generated essays. The results reveal a weak positive correlation ($\beta = 2.659^{*}; p <$

0.1) between the number of AI essays evaluated and the probability of correct identification, suggesting that as psychologists evaluate more AI-generated essays, they may become slightly better at distinguishing them from human-written content. This supports the conclusion that while human evaluation is complex, experience can lead to improved accuracy in assessing nuanced cognitive processes.

	Dependent variable:	
	Total Correct	
	(1)	(2)
Total AI	4.329***	2.659*
	(0.225)	(1.387)
Total Pass		-0.162
		(0.185)
Total Grade		0.385
		(0.278)
Constant	-2.029	-5.272
	(4.645)	(5.235)
Observations	15	15
R^2	0.966	0.972
Adjusted R^2	0.964	0.964
Residual Std. Error	8.319 (df = 13)	8.213 (df = 11)
F Statistic	371.322*** (df = 1; 13)	127.781*** (df = 3; 11)
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 15: Hedging Effects on Psychologist Graders

Additionally, we observed that an essay evaluated as written by AI is more likely to be graded as failing or demonstrating stereotyped thinking. However, the effect size of this observation is relatively small (odds ratio = -1.049; $p < 0.001$). This finding further confirms that while there is some correlation between the ability to identify AI-generated content and the grade outcomes, it is not a substantial factor driving the overall evaluation of critical thinking.

These results confirm that while essay length and writing quality correlate with critical thinking, they are not sufficient indicators on their own. The robustness of our findings lies in nuanced evaluations by human psychologists who can detect deeper cognitive processes. The consistency of grading, despite varying familiarity with AI-generated content, further validates our methodology and highlights the

importance of human judgment in assessing complex cognitive skills.

C.3.2 Aligning LLMs with Human Experts

In this section, we further explore the replacement of human expert scores with LLMs. First, we test whether the GPT-3 level LLM, which we used to incentivize our measure of critical thinking, can replace psychologist scores to identify critical thinking in each essay. When we replace psychologist grades with LLM-generated scores as evidence of critical thinking, differences in the proportion of changing thinking styles are absent for participants in all storytelling contexts, as shown in Table 16. In other words, this baseline GPT-3 LLM does not accurately detect signals of critical thinking.

Treatment	*Newspaper*	*Twitter*	*Biased Twitter*	*Facebook*
Newspaper	·	·	·	·
Twitter	−0.755 (0.055)	·	·	·
Biased Twitter	−0.867 (0.056)	−0.125 (0.055)	·	·
Facebook	−1.326 (0.056)	−0.590 (0.054)	−0.457 (0.055)	·
N	170	190	184	181

Standard errors in parentheses
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 16: *Z-Scores* with LLM (GPT-3 level) Grading System

Second, we use a more advanced LLM, *ChatGPT-4+* from OpenAI, to investigate whether the current LLM techniques are more robust baseline technologies to detect signals of critical thinking accurately. When we proceed with simple text detection, 44.87% of the grades do not match the psychologist majority grades and the LLM grades. Furthermore, when providing the LLM with the same instructions that the psychologists received, 48.01% of the LLM grades still do not match the psychologists' grades. Provided with half of the psychologists' grades and digital essays, *ChatGPT-4+* predicted incorrect grades for 29.23% of 366 observations.

To further support our findings, additional analyzes reveal that the quality of the essay (word count and grammar) correlates with higher critical thinking assessments by psychologists, highlighting a potential confounding factor. Table 17 shows that essays that show higher word counts and better grammar have higher chances of receiving pass grades or being recognized as demonstrating critical thinking.

	Dependent variable:
	Ex Post
Digital Words	0.001^{***}
	(0.0005)
Grammarly Digital Grade	0.005^{***}
	(0.002)
Constant	-0.249^{*}
	(0.145)
Observations	588
R^2	0.027
Adjusted R^2	0.024
Residual Std. Error	0.461 (df = 585)
F Statistic	8.142^{***} (df = 2; 585)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 17: Regression Estimate of Psychologist Grades by Essay Qualities

Furthermore, when examining participants with a high need for cognition using LLM grades, we found no significant differences in the changes between thinking styles in storytelling contexts, contradicting our main results. Table 18 supplements Table 16 by exploring the effects of storytelling contexts on thinking styles among participants with a high need for cognition based on LLM grades. When we replace psychologist grades with LLM grades as evidence of critical thinking, differences in changing reasoning styles across storytelling contexts are absent for agents with a high need for cognition. Like Table 16, we observe that LLMs cannot replace psychology grades, as Table 18 presents results inconsistent with our main finding.

Treatment	*Newspaper*	*Twitter*	*Biased Twitter*	*Facebook*
Newspaper	·	·	·	·
Twitter	−1.567	·	·	·
	(0.086)			
Biased Twitter	−1.735*	−0.170	·	·
	(0.086)	(0.085)		
Facebook	−1.720*	−0.167	0.003	·
	(0.086)	(0.086)	(0.086)	
N	71	77	81	72

Standard errors in parentheses
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 18: Z-Scores for LLM Grades for High Need for Cognition Participants

These insights underscore the limitations of current LLM techniques in assessing critical thinking and emphasize the continued importance of expert human evaluation in this domain. Our findings suggest that while LLMs, such as GPT-3 and ChatGPT-4+, can provide some insight, they lack the nuanced understanding that human experts bring to the evaluation of critical thinking. This highlights the need for ongoing development in AI techniques to better align with human cognitive assessments and the necessity of maintaining human oversight in critical evaluations.

D Conceptual Framework with Three Thinking Stages

We propose an additional model in which agents can be in three different states of critical thinking: not engaged with critical thinking, performing critical thinking (either in its first or second state), and having finished performing critical thinking. In our two-stage model, we considered performing critical thinking to be having finished performing it. In this scenario, we propose a three-stage (not fully identified) model that considers the three stages distinctively.

In this economy, the object of interest is the distribution of reasoned preferences over a binary policy space in a large population. A three-dimensional type characterizes each j inside the population.

$$(x_j, y_j, i_j) \in \mathcal{J} := \{0,1\} \times \{0,1\} \times \{0,1\}$$

Where x_j represents the naive preference individual j would self-report when presented with a dilemma for the first time – that is, by definition, before undergoing a critical thinking phase; y_j differs potentially from x_j as it represents the reasoned preference that j holds after completing their period of critical thinking; the cogni-

tive type i_j refers to the cognitive type i_j, interacting with the format, determines how easily individual j moves into (and out of) critical thinking.

Individuals go through a three-step process of "critical thinking" as they form their preferences. The process begins with a "naive-self" state, followed by a period of critical thinking, and ultimately leading to a "reasoned preferences-self" state. We assume that this process is irreversible and that once individuals reach a reasoned preferences-self state, they no longer question their preferences. There is no additional "information" that has to come and change their worldview: the process of critical thinking provides a final and reasoned preference answer to dilemmas. When asked to report their preferences on a policy issue, individuals in either their naive self or reasoned preferences self-state will vote according to their respective preferences, x_j, y_j, respectively. Those who are still in the critical thinking phase will abstain from voting.

The transition between the different mental models is determined by an individual's cognitive style and the characteristics of the storytelling context. Hence, the storytelling context is instrumental in the agent's transition from a naive state to a reasoned preferences one. By constructing our model, this transition is captured by the critical thinking phase. An economy of reasoned preferences is preferable from efficiency and welfare perspectives to an economy of naive preferences. We formally present such an economy below.

D.1 Model Identification

Using reported preferences of individuals that do the $S(Stereotype) \rightarrow T(Type)$ transition (i.e. we observe ex ante x_S then y), we get

$$\mathbb{E}\left[x_S \,|y = 1\right] = (1 - \beta) + \beta p_S$$

$$\mathbb{E}\left[x_S \,|y = 0\right] = \beta p_S$$

which gives the estimators

$$\hat{\beta} = 1 - \left(\bar{x}_{S|1} - \bar{x}_{S|0}\right)$$

and

$$\hat{p}_S = \frac{\bar{x}_{S|0}}{\hat{\beta}}$$

clearly $\hat{p} = \bar{y}$. Finally, using the reported preferences of individuals that do the $A \rightarrow T(Type)$ transition we can estimate ξ_A as

$$\mathbb{E}\left[x_A \,|y = 1\right] = \xi_A$$

$$\mathbb{E}\left[x_A \,|y=0\right] = 1 - \xi_A$$

so $\hat{\xi}_A = \bar{x}_{NU|1}$ or $\hat{\hat{\xi}}_A = 1 - \bar{x}_{NU|0}$. Notice that we can test the assumed symmetry by testing that $\hat{\xi}_A = \hat{\hat{\xi}}_A$. Since our dataset has few agents that start in A, this test has almost no power.

D.2 General Results

The basic decomposition

$$\begin{aligned} W_E &= W_P + Bias \\ -\mathbb{E}\left[(p-\bar{p})^2\right] &= -\left(\mathbb{E}\left[(p-\hat{p})^2\right] + \mathbb{E}\left[(\hat{p}-\bar{p})^2\right]\right) \end{aligned}$$

is still clearly valid. However, $\bar{p}$ is now given by

$$\begin{aligned} \bar{p} &= \mu_T p + \mu_S\left(\beta p_S + (1-\beta)\, p\right) + \mu_A\left[1 - p + \xi_A\left(2p-1\right)\right] \\ &= \alpha_0\left(t,\lambda\right) + \alpha_1\left(t,\lambda\right) p + \alpha_2\left(t,\lambda\right) p_S \end{aligned}$$

with

$$\begin{aligned} \alpha_0 &= \mu_A\left(1-\xi_A\right) \\ \alpha_1 &= 1 - \beta\mu_S - 2\mu_A\left(1-\xi_A\right) \\ \alpha_2 &= \beta\mu_S \end{aligned}$$

where $\hat{p}$ (that was wrong in the previous file since for non-normal random variables, we do not know the expectation of p given the convex combination $\beta p_s + (1-\beta)\, p$) is given by

$$\hat{p} = \frac{\frac{\bar{p} - \left[\alpha_0 + \alpha_2\mu_y\right]}{\alpha_1}\alpha_1^2\sigma_x^2 + \alpha_2^2\sigma_y^2\mu_x}{\alpha_1^2\sigma_x^2 + \alpha_2^2\sigma_y^2} = \tilde{\alpha_0}\left(t,\lambda\right) + \tilde{\alpha_1}\left(t,\lambda\right) p + \tilde{\alpha_2}\left(t,\lambda\right) p_S$$

$$\frac{\frac{\alpha_0(t,\lambda) + \alpha_1(t,\lambda)p + \alpha_2(t,\lambda)p_S - \left[\alpha_0 + \alpha_2\mu_y\right]}{\alpha_1}\alpha_1^2\sigma_x^2 + \alpha_2^2\sigma_y^2\mu_x}{\alpha_1^2\sigma_x^2 + \alpha_2^2\sigma_y^2}$$

so

$$\tilde{\alpha_1}\left(t,\lambda\right) = \frac{\alpha_1^2\sigma_x^2}{\alpha_1^2\sigma_x^2 + \alpha_2^2\sigma_y^2}$$

$$\tilde{\alpha_2}(t,\lambda) = \frac{\alpha_2\alpha_1\sigma_x^2}{\alpha_1^2\sigma_x^2 + \alpha_2^2\sigma_y^2}$$

when is it

$$\tilde{\alpha_1}(t,\lambda) = \alpha_1(t,\lambda) \iff \frac{\alpha_1^2\sigma_x^2}{\alpha_1^2\sigma_x^2 + \alpha_2^2\sigma_y^2} = \alpha_1 \iff$$
$$\text{Same } \sigma = \alpha_1(1-\alpha_1) = \alpha_2^2 \iff (1-\beta\mu_S)(\beta\mu_S) = (\beta\mu_S)^2$$

It is immediate to check that $\tilde{\alpha_1}(t,\lambda) = \alpha_1(t,\lambda) \iff \tilde{\alpha_2}(t,\lambda) = \alpha_2(t,\lambda)$ so we actually have a single equation, and we need to claim that $\exists$ time such that

$$\frac{\alpha_1^2\sigma_x^2}{\alpha_1^2\sigma_x^2 + \alpha_2^2\sigma_y^2} = \alpha_1 \iff \frac{(1-\beta\mu_S-2\mu_A(1-\xi_A))^2\sigma_x^2}{(1-\beta\mu_S-2\mu_A(1-\xi_A))^2\sigma_x^2 + (\beta\mu_S)^2\sigma_y^2} = (1-\beta\mu_S-2\mu_A(1-\xi_A))$$
$$\iff (1-\beta\mu_S-2\mu_A(1-\xi_A))\sigma_x^2 = (1-\beta\mu_S-2\mu_A(1-\xi_A))^2\sigma_x^2 + (\beta\mu_S)^2\sigma_y^2$$
$$\iff (1-\beta\mu_S-2\mu_A(1-\xi_A))\sigma_x^2(\beta\mu_S+2\mu_A(1-\xi_A)) = (\beta\mu_S)^2\sigma_y^2$$
$$\iff \frac{(1-\beta\mu_S-2\mu_A(1-\xi_A))(\beta\mu_S+2\mu_A(1-\xi_A))}{(\beta\mu_S)^2} = \frac{\sigma_y^2}{\sigma_x^2}$$

now substituting μ_S, μ_A we have the LHS is increasing to ∞ in t, therefore there is a unique solution provided that it starts below $\frac{\sigma_y^2}{\sigma_x^2}$, that is if $\frac{1-\beta}{\beta} < \frac{\sigma_y^2}{\sigma_x^2}$ (β is large enough)

as $\mu_S \to 0$, this explodes (there is always bias in the limit), while at the beginning, there is zero bias iff

$$\frac{1-\beta}{\beta} = \frac{\sigma_y^2}{\sigma_x^2}$$

[example, $\sigma_x^2 = \frac{1}{2}, \sigma_y^2 = \frac{1}{6}$ $\beta = \frac{3}{4} \implies \frac{1-\beta}{\beta} = \frac{1}{3}$]

This is the zero-bias time. Even more special cases $\xi_A = 1$

$$\frac{(1-\beta\mu_S)^2\sigma_x^2}{(1-\beta\mu_S)^2\sigma_x^2 + (\beta\mu_S)^2\sigma_y^2} = (1-\beta\mu_S)$$

$$(1-\beta\mu_S)\sigma_x^2 = (1-\beta\mu_S)^2\sigma_x^2 + (\beta\mu_S)^2\sigma_y^2$$

$$\left(\frac{1-\beta\mu_S}{\beta\mu_S}\right) = \frac{\sigma_y^2}{\sigma_x^2}$$

Since the LHS is decreasing in μ_S and the RHS is increasing, then there is at most one solution. It has no if

$$\frac{1-\beta}{\beta} > \frac{\sigma_y^2}{\sigma_x^2}$$

Furthermore, we get

$$W_P = -\mathbb{E}\left[(p-\hat{p})^2\right] = -\frac{\alpha_2^2\sigma_y^2\sigma_x^2}{\alpha_1^2\sigma_x^2+\alpha_2^2\sigma_y^2}$$

$$\text{If } \sigma_x^2 = \sigma_y^2 = \sigma^2 = -\frac{\alpha_2^2}{\alpha_1^2+\alpha_2^2}\sigma^2$$

$$\xi_A = 1 = \frac{(\beta\mu_S)^2}{2\beta\mu_S\left[1-\beta\mu_S\right]+1}\sigma^2??$$

D.2.1 Aside: No Bias

The condition for no bias is that coefficients in $\bar{p}$ are the same as in $\hat{p}$ that is,

$$If \exists t : B(t) = 0$$

$$\alpha_0 + \alpha_1 p + \alpha_2 p_S$$

$$\frac{\frac{\alpha_1 p+\alpha_2 p_S-\alpha_2\mu_y}{\alpha_1}\alpha_1^2\sigma_x^2+\alpha_2^2\sigma_y^2\mu_x}{\alpha_1^2\sigma_x^2+\alpha_2^2\sigma_y^2} = \frac{\alpha_2^2\sigma_y^2\mu_x-\alpha_2\alpha_1\sigma_x^2\mu_y}{\alpha_1^2\sigma_x^2+\alpha_2^2\sigma_y^2} + \frac{\alpha_1^2\sigma_x^2}{\alpha_1^2\sigma_x^2+\alpha_2^2\sigma_y^2}p + \frac{\alpha_2\alpha_1\sigma_x^2}{\alpha_1^2\sigma_x^2+\alpha_2^2\sigma_y^2}p_S$$

$$\alpha_0 = \frac{\alpha_2^2\sigma_y^2\mu_x-\alpha_2\alpha_1\sigma_x^2\mu_y}{\alpha_1^2\sigma_x^2+\alpha_2^2\sigma_y^2} = \mu\frac{\alpha_2^2\sigma_y^2-\alpha_2\alpha_1\sigma_x^2}{\alpha_1^2\sigma_x^2+\alpha_2^2\sigma_y^2}$$

$$\alpha_1 = \frac{\alpha_1^2\sigma_x^2}{\alpha_1^2\sigma_x^2+\alpha_2^2\sigma_y^2}$$

$$\alpha_2 = \frac{\alpha_2\alpha_1\sigma_x^2}{\alpha_1^2\sigma_x^2+\alpha_2^2\sigma_y^2}$$

notice that if $\beta = \frac{1}{2}$ then at $t = 0$ we have a solution iff σ are the same at $t = 0$,

$$\alpha_0 = 0$$

$$\alpha_1 = \frac{1}{2}$$

$$\alpha_2 = \frac{1}{2}$$

$$\alpha_0 = \frac{\alpha_2^2\sigma_y^2\mu_x-\alpha_2\alpha_1\sigma_x^2}{\alpha_1^2\sigma_x^2+\alpha_2^2\sigma_y^2}$$

$$\alpha_1 = \frac{\alpha_1^2\sigma_x^2}{\alpha_1^2\sigma_x^2+\alpha_2^2\sigma_y^2}$$

$$\alpha_2 = \frac{\alpha_2\alpha_1\sigma_x^2}{\alpha_1^2\sigma_x^2+\alpha_2^2\sigma_y^2}$$

and

$$\mathbb{E}\left[\hat{p}|p\right] = \frac{\mathbb{E}\left[\frac{\alpha_1 p + \alpha_2 p_S - \alpha_2 \mu_y}{\alpha_1}\right] \alpha_1^2 \sigma_x^2 + \alpha_2^2 \sigma_y^2 \mu_x}{\alpha_1^2 \sigma_x^2 + \alpha_2^2 \sigma_y^2} = \frac{\alpha_1^2 \sigma_x^2 p + \alpha_2^2 \sigma_y^2 \mu_x}{\alpha_1^2 \sigma_x^2 + \alpha_2^2 \sigma_y^2}$$

and

$$\mathbb{E}\left[\hat{p}\right] = \frac{\alpha_1^2 \sigma_x^2 \mu_x + \alpha_2^2 \sigma_y^2 \mu_x}{\alpha_1^2 \sigma_x^2 + \alpha_2^2 \sigma_y^2} = \mu_x$$

D.2.2 Positive and Normative Welfare Expressions

The general formula is in the mathematical file, under the restriction $\mu_x = \mu_y$ and $\sigma_x = \sigma_y$ we get

$$W_E = -\left[\left(\alpha_0 - \left(1 - \alpha_1 - \alpha_2\right)\mu\right)^2 + \left(\left(1 - \alpha_1\right)^2 + \alpha_2^2\right)\sigma^2\right]$$

We have welfare at $t = 0$, where $\mu_S = 1$. Namely

$$W_E = -\beta^2 \left(\underbrace{\left(\mu_x - \mu_y\right)^2}_{\text{Prior Bias}} + \sigma_x^2 + \sigma_y^2 \right)$$

$$W_P = -\beta^2 \frac{\sigma_x^2 \sigma_y^2}{\sigma_x^2 \left(1 - \beta\right)^2 + \sigma_y^2 \beta^2}$$

Then,

$$\frac{W_E}{W_P} = \frac{\left(\mu_x - \mu_y\right)^2 + \sigma_x^2 + \sigma_y^2}{\frac{\sigma_x^2 \sigma_y^2}{\sigma_x^2 (1-\beta)^2 + \sigma_y^2 \beta^2}}$$

$$\text{Assume equal } \sigma = \frac{\left(\mu_x - \mu_y\right)^2 + 2\sigma^2}{\frac{\sigma^2}{(1-2\beta+2\beta^2)}} = \underbrace{\frac{\left(\mu_x - \mu_y\right)^2}{\frac{\sigma^2}{(1-2\beta+2\beta^2)}}}_{>0} + 2\left(1 - 2\beta + 2\beta^2\right)^2 > 2\left(\frac{1}{2}\right) = 1$$

so if $\mu_x = \mu_y$ (no prior bias), then $W_E\left(0\right) = W_P\left(0\right)$ iff $\sigma_x = \sigma_y$.

Results

$W_E > W_P$ this is because the bias/variance decomposition

$$\begin{aligned} W &= -\mathbb{E}\left[(p-\bar{p})^2\right] = -\mathbb{E}\left[\left(\left(1-\mu_T-\mu_A\left[2\xi_A-1\right]-\mu_S\left(1-\beta\right)\right)p+\beta\mu_S p_S+\mu_A\left(1-\xi_A\right)\right)^2\right] \\ &= -\mathbb{E}\left[(p-\hat{p}+\hat{p}-\bar{p})^2\right] = -\left(\mathbb{E}\left[(p-\hat{p})^2\right]+\mathbb{E}\left[(\hat{p}-\bar{p})^2\right]+\cancel{2\mathbb{E}\left[(p-\hat{p})(\hat{p}-\bar{p})\right]}\right) \\ &= -\left(\underbrace{\mathbb{E}\left[(p-\hat{p})^2\right]}_{\text{Precision of election}}+\underbrace{\mathbb{E}\left[(\hat{p}-\bar{p})^2\right]}_{\text{Bias of elections}}\right) \end{aligned}$$

finally holds, the election has a bias.

The full characterization of the derivative (assuming equal μ and σ)

$$\frac{\mathrm{d}}{\mathrm{d}t}W_E|_{t=0} = -4\beta\lambda_1\sigma^2\left(1-\beta-\xi_A\right)$$

Instead assuming only equal μ we have

$$\frac{\mathrm{d}}{\mathrm{d}t}W_E|_{t=0} = -2\beta\lambda_1\left(\beta\sigma_y^2-\sigma_x^2\left(2\left(1-\xi_A\right)-\beta\right)\right)$$

so

$$\frac{\mathrm{d}}{\mathrm{d}t}W_E|_{t=0} > 0 \iff 1-\beta < \xi_A$$

or in general

$$\frac{\beta}{2\left(1-\xi_A\right)-\beta} < \frac{\sigma_x^2}{\sigma_y^2}$$

a sensible condition. Also, λ_1 magnifies either the positive or the negative change local to 0 and in particular if $1-\beta > \xi_A$ then more λ_1 is bad for welfare local to $t=0$. On the contrary,

$$\frac{\mathrm{d}}{\mathrm{d}t}W_P|_{t=0} = \frac{2\left(1-\beta\right)\beta^2\lambda_1\sigma^2\left(2\xi_A-1\right)}{\mathrm{sthg}^2} > 0$$

and the welfare of the unconstrained principal is [, but this is just a conjecture not falsified by Math plots] always increasing in both λ_1, t.

Conjecture

W_P is increasing in t (and λ_1)— We show that

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}W_P &\propto -\left[\underbrace{2\left(1-\xi_A\right)\mu_S}_{+}\underbrace{\frac{\mathrm{d}}{\mathrm{d}t}\mu_A}_{?}+\underbrace{\left(1-2\left(1-\xi_A\right)\mu_A\right)}_{+}\underbrace{\frac{\mathrm{d}}{\mathrm{d}t}\mu_S}_{-}\right]\\ &=\underbrace{\exp\left\{-\left(\lambda_1+\lambda_2\right)t\right\}}_{+}\lambda_1\left(2\left(1-\xi_A\right)-\exp\left\{\lambda_2 t\right\}\right)2\left(1-\xi_A\right)-\exp\left\{\lambda_2 t\right\}<2\left(1-\xi_A\right)-1\\ &=1-2\xi_A<0\end{aligned}$$

when computed in

$$\frac{\mathrm{d}}{\mathrm{d}\lambda_1}W_P \propto -\left[\underbrace{2\left(1-\xi_A\right)\mu_S}_{+}\underbrace{\frac{\mathrm{d}}{\mathrm{d}\lambda_1}\mu_A}_{?}+\underbrace{\left(1-2\left(1-\xi_A\right)\mu_A\right)}_{+}\underbrace{\frac{\mathrm{d}}{\mathrm{d}\lambda_1}\mu_S}_{-}\right]$$

which has the same sign as

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\lambda_1}W_P \propto\ &-\exp\left\{\lambda_2 t\right\}\lambda_2^2 t\\ &+2\lambda_2\left[\exp\left\{\lambda_2 t\right\}\lambda_1 t+\left(1-\xi_A\right)\exp\left\{\left(\lambda_2-\lambda_1\right)t\right\}\right.\\ &\left.-(1-\xi_A)(1+\lambda_1 t)\right]\\ &-\lambda_1^2 t\left(\exp\left\{\lambda_2 t\right\}-2(1-\xi_A)\right)\end{aligned}$$

Furthermore, ξ_A

$$\begin{aligned}&=-\exp\left\{\lambda_2 t\right\}\lambda_2^2 t+2\lambda_2\left[\left(\exp\left\{\lambda_2 t\right\}\lambda_1 t\right)\right]-\lambda_1^2 t\left(\exp\left\{\lambda_2 t\right\}\right)\\ &=-t\exp\left\{\lambda_2 t\right\}\left(\lambda_2-\lambda_1\right)^2+2\lambda_2\left(1-\xi_A\right)\left[\exp\left\{\left(\lambda_2-\lambda_1\right)t\right\}-\left(1+\lambda_1 t\right)\right]+2\lambda_1^2 t\left(1-\xi_A\right)\\ &=\underbrace{-t\exp\left\{\lambda_2 t\right\}\left(\lambda_2-\lambda_1\right)^2}_{negative}+2\lambda_2\left(1-\xi_A\right)\left[\exp\left\{\left(\lambda_2-\lambda_1\right)t\right\}-\left(1+\lambda_1 t\right)+2\lambda_1^2 t\right]\end{aligned}$$

Now, if the second supplement is negative, then we are done, so assume it is positive; that is

$$\exp\left\{\left(\lambda_2-\lambda_1\right)t\right\}-\left(1+\lambda_1 t\right)+2\lambda_1^2 t>0$$

then the sum is smaller than

$$\underbrace{-t\exp\left\{\lambda_2 t\right\}\left(\lambda_2-\lambda_1\right)^2}_{<0}+\lambda_2\left[\exp\left\{\left(\lambda_2-\lambda_1\right)t\right\}\left(1+\lambda_1 t\right)+2\lambda_1^2 t\right]$$
$$\begin{aligned}&=\lambda_1^2 t-\exp\left\{\lambda_2 t\right\}\left(\lambda_2-\lambda_1\right)^2 t+\lambda_2\left(-1+\exp\left\{\left(\lambda_2-\lambda_1\right)t\right\}-\lambda_1 t\right)\\&=\lambda_1^2 t+\lambda_2\exp\left\{\left(\lambda_2-\lambda_1\right)t\right\}-\left[\exp\left\{\lambda_2 t\right\}\left(\lambda_2-\lambda_1\right)^2 t+\lambda_2\left(1+\lambda_1 t\right)\right]\\&<0\end{aligned}$$

so it remains to show that this is always negative; if $\lambda_1 \approx 0$

$$-\lambda_2\left(1-\exp\left\{\lambda_2 t\right\}\left(1-\lambda_2 t\right)\right)<-\lambda_2^2 t<0$$

W_E has interesting comparative statistics due to the interaction with bias. In particular, it seems that for $\beta >$ stgh, then [if there is no prior bias, $\mu_x = \mu_y$] there is a time t such that $Bias\left(t\right)=0$ because the evolution of μ_S, μ_A is such that $\alpha^E = \alpha^P$. This seems interesting, possibly a result of putting in a proposition.

D.2.3 Novel Results: Non-Monotonicity, Compensation Effect, and Costly Voting

Based on our model, we can draw three main results and one additional interesting result.

Inefficiency of Twitter economy and non-monotonicity in election times. The first result relates to the political institutions of the digital economy. From our model, a Twitter-Facebook economy where everyone can speak their mind is not necessarily good. In fact, we want only those who went through critical thinking to vote. Following this point, what follows from the naturally occurring question *when do we want to hold elections*? Our model clearly implies non-monotonicity in time for election periods.

Typology of voting-users and adverse selection. The second result relates to the typology of voting users. The *"clients" of news outlets*, in a microfoundation of the λ functions, are either low i partisans (which look at it for fun) or frustrated critical thinking voting-users that look for some facts (positive predictions). On a related but different point, we can identify the *adverse selection in the vote-force* (under some conditions, the strengths of the naive pool weaken) and how the storytelling type amplifies or reduces this issue (always true that it is better if only types vote, at least

in the symmetric case).

Partisan format and compensation effect. The third and most intriguing result relates to the storytelling type. We can study *impact of different storytelling contexts* (more in-depth, helps the high i, but how it correlates with α): more in-depth, with a somewhat primitive nature. In particular and more interestingly, we can allow for *asymmetries*: either there is A "better" policy (say $\beta = 1$, so upon reflecting, everyone agrees 1 is right) or stereotypes of one side are less likely to enter critical thinking (evidence that conservatives are overconfident), how does this change the outcome, as well as the incentives for the critical thinking agents (that may vote for those that are less confident because of the bias in the type pool). The problem of asymmetries is that a partisan format, or the fact that one stereotype is more attractive than the other, makes the problem of agents in critical thinking more problematic: remember they are smart but unwise, so they cannot ignore the fact of a stereotyped partisan pool, either because stereotypes are more resistant, or because the shifts the stereotypes. Hence, we propose to explain such a situation by an effect that we label the "Compensation Effect": When you perceive the device to be partisan in one direction, you vote in the opposite direction when in critical thinking.

The benefits of making voting costly. A resulting and potentially controversial consequence of such an asymmetry is that *voting costs* in this situation may be positive because they can also exclude the strategic types that recognize the naive pool is partisan and cannot morally abstain or vote against their type. They can use the excuse not to vote.

D.3 Proofs

E Experimental Design with Three mental models

In the experiment, we collected data to decompose the critical thinking process into three stages: N, U, C. Here, N remains unchanged. U denotes an intermediate transitory stage during which agents experience internal uncertainty about the formation of their reasoned preferences. Finally, now C denotes the stage in which the agents have completed the critical thinking process and formed their reasoned preferences.

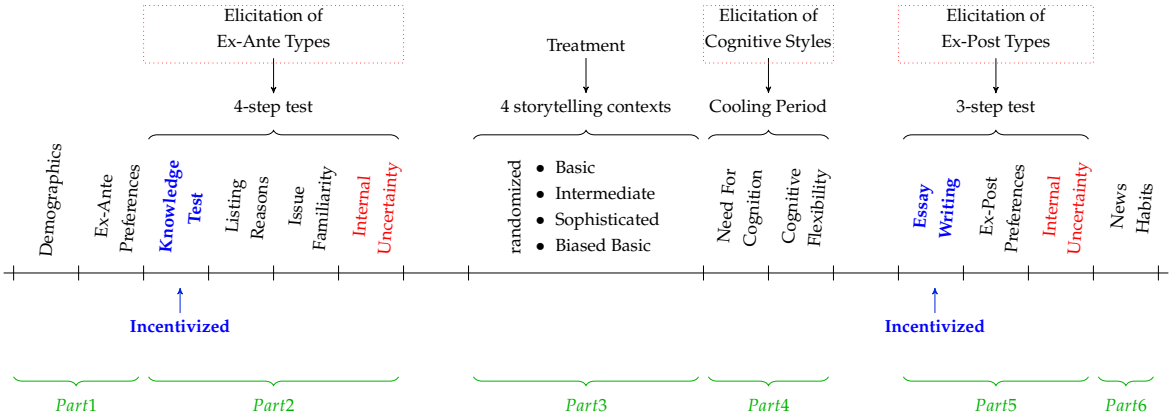


Table 19 shows the classification strategy of participants as *S*tereotype, *A*ware, and *T*ype.

Treatment	T	A	S
BEFORE	Knowledge Test Score $> \tau_{KTS}$ Issue Familiarity = 1 **Internal Uncertainty** $\neq 0$ Reasons List $> \tau_{RL}$	Knowledge Test Score $> \tau_{KTS}$ Issue Familiarity = 1	
AFTER	Psychologists Grade = Pass	Else	

Table 19: CLASSIFICATION **STRATEGY** BEFORE/AFTER TREATMENT

The analysis presents the frequencies of the three states of participants before and after the treatment.

	S_1	A_1	T_1
S_0	457	150	38
A_0	27	7	2
T_0	24	3	7

Table 20: TABLE: FREQUENCIES BEFORE/AFTER TREATMENT

Table 21 collects Z-Scores for differences in proportions of thinking style changes when mental models consist of S, A, and T.

Treatment	*Newspaper*	*Twitter*	*Biased Twitter*	*Facebook*
Newspaper	·	·	·	·
Twitter	0.262 (0.048)	·	·	·
Biased Twitter	−0.504 (0.050)	−0.793 (0.048)	·	·
Facebook	−1.500 (0.051)	−1.834* (0.049)	−1.003 (0.051)	·
N	170	190	184	181

Standard errors in parentheses
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 21: Z-Scores for Three States